REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

AD-A267 197

REPORT DATE	3. REPORT TYPE AND DATES COVERED
	ANNUAL 1 Jun 92 TO 31 May 93

4. TITLE AND SUBTITLE

DEVELOPMENT OF PATTERN RECOGNITION TECHNIQUES FOR THE EVALUATION OF TOXICANT IMPACTS TO MULTISPECIES SYSTEMS

5. FUNDING NUMBERS

AFOSR-91-0291
61103D
3484
RS

6. AUTHOR(S)

Dr Wayne G. Landis

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute of Environmental Toxicilogy & Chemistry
Huxley College of Environmental Studies
Western Washington University
Bellingham, WA 98225

8. PERFORMING ORGANIZATION REPORT NUMBER

AFOSR-TR- 93 0[illegible]

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR/NL
110 Duncan Avenue, Suite B115
Bolling AFB DC 20332-0001

Dr Walter Kozumbo

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release;
distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

A common assumption in environmental toxicology is that after the initial impact, ecosystems recover to resemble the control state. This assumption may be based more on our inability to observe an ecosystem with sufficient resolution to detect differences, than reality. Recent findings of complex and perhaps chaotic dynamics in two relatively simple types of microcosms demonstrate that complex dynamics and non-equilibrium systems are the rule rather than the exception. In the Standardized Aquatic Microcosm and the Mixed Flask Culture (MFC) microcosms, multivariate analysis and clustering methods derived from artificial intelligence research was able to differentiate oscillations that separate the treatments from the reference group, followed by what would normally appear as recovery, followed by another separation into treatment groups as distinct from the reference treatment. The explanation may be that the oscillations are the result of the intrinsic chaotic behavior of population interactions, of which the alteration of detrital quality is but one of many. In fact, preliminary data indicate that material derived from the jet fuel may be released back into the water column due to the decay or organic material. The initial impact of the toxicant re-set the dosed communities into different regions of the

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
(U)	(U)	(U)	(UL)

n-dimensional space where recovery may be an illusion due to the incidental overlap of the oscillation trajectories occurring along a few axes. We now use the new visualization technique of space-time worms to see the trajectories of the ecosystems through n-dimensional ecosystem space. The dynamics appear to have little regularity and resemble chaotic systems in the lack of repeatability and the importance of initial conditions. The dynamics of ecosystems may be more closely related in terms of basic dynamics to such phenomena as turbulence and weather formation. The implications for risk assessment and resource management are being examined.

Development Of Pattern Recognition Techniques for the Evaluation of Toxicant Impacts to Multispecies Systems

USAFOSR
Grant No. AFOSR-91-0291 DEF

Wayne G. Landis
Robin A. Matthews

Institute of Environmental Toxicology and Chemistry, Huxley College of Environmental Studies, Western Washington University, Bellingham, WA 98225, 206/650-6136

Geoffrey B. Matthews

Computer Science Department, Western Washington University, Bellingham, WA 98225, 206/650-3797

18 June 1993
Annual Technical Report for Period 1 June 1992 - 31 May 1993

UNCLASSIFIED

Prepared for
LIFE SCIENCES DIRECTORATE
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Department of the Air Force
Bolling Air Force Base, DC 20332-6448

Development Of Pattern Recognition Techniques for the Evaluation of Toxicant Impacts to Multispecies Systems

USAFOSR
Grant No. AFOSR-91-0291 DEF

Wayne G. Landis
Robin A. Matthews

Institute of Environmental Toxicology and Chemistry, Huxley College of Environmental Studies, Western Washington University, Bellingham, WA 98225, 206/650-6136

Geoffrey B. Matthews

Computer Science Department, Western Washington University, Bellingham, WA 98225, 206/650-3797

18 June 1993
Annual Technical Report for Period 1 June 1992 - 31 May 1993

UNCLASSIFIED

Prepared for
LIFE SCIENCES DIRECTORATE
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Department of the Air Force
Bolling Air Force Base, DC 20332-6448

Program Summary 92-93

A common assumption in environmental toxicology is that after the initial impact, ecosystems recover to resemble the control state. This assumption may be based more on our inability to observe an ecosystem with sufficient resolution to detect differences, than reality. Recent findings of complex and perhaps chaotic dynamics in two relatively simple types of microcosms demonstrate that complex dynamics and non-equilibrium systems are the rule rather than the exception.

In the Standardized Aquatic Microcosm and the Mixed Flask Culture (MFC) microcosms, multivariate analysis and clustering methods derived from artificial intelligence research was able to differentiate oscillations that separate the treatments from the reference group, followed by what would normally appear as recovery, followed by another separation into treatment groups as distinct from the reference treatment. The explanation may be that the oscillations are the result of the intrinsic chaotic behavior of population interactions, of which the alteration of detrital quality is but one of many. In fact, preliminary data indicate that material derived from the jet fuel may be released back into the water column due to the decay or organic material. The initial impact of the toxicant re-set the dosed communities into different regions of the n-dimensional space where recovery may be an illusion due to the incidental overlap of the oscillation trajectories occurring along a few axes.

We now use the new visualization technique of space-time worms to see the trajectories of the ecosystems through n-dimensional ecosystem space. The dynamics appear to have little regularity and resemble chaotic systems in the lack of repeatability and the importance of initial conditions. The dynamics of ecosystems may be more closely related in terms of basic dynamics to such phenomena as turbulence and weather formation. The implications for risk assessment and resource management are being examined.

Program Objectives

The principal objective of this project is to examine the patterns in toxicity data from experiments using two microcosm protocols. We use nonmetric clustering, a multivariate pattern recognition technique developed by Matthews and Hearne (1991), for our primary pattern analyses. NMC has been shown to work well on a variety of ecological data sets (Matthews and Hearne, 1991). The results from the NMC analyses are then compared with those from other standard multivariate techniques to compare the utility of each technique for analyzing aquatic toxicity data.

Specific objectives are:

- Conduct one series of toxicity tests using the SAM and Mixed Flask Culture (MFC) protocols with 3 complex toxicants such as the water soluble fraction of JP-4, shale derived JP-4, and JP-8.
- For at least one of the complex toxicants, conduct a second complete series of toxicity tests (SAM and MFC) to compare similarities between parallel tests.

- Examine the SAM and MFC complex toxicant data using NMC, linear discriminant analysis, correspondence analysis, and metric clustering (k-means using Euclidean and cosine distances).

- Examine existing SAM data from experiments conducted previously for copper sulfate, brass, and graphite using NMC, linear discriminant analysis, correspondence analysis, and metric clustering.

- Describe a protocol that can be used for analyzing multispecies toxicity data. This protocol will incorporate a discussion of the advantages and limitations of the different multivariate analytical tools that were tested during this project.

Status of the Research

The results from the first and second years of the research program have been presented at the 1992 Annual Meeting of the Society for Environmental Toxicology and Chemistry (SETAC) in Cincinnati, the 1993 First SETAC World Congress in Lisbon, Portugal, and the recent Third Annual Symposium for Environmental Toxicology and Risk Assessment sponsored by Committee E47 of the American Society for Testing and Materials (ASTM) in Atlanta. In addition to these presentations, we have also presented our research results during several invited seminars, including the Keynote Address , "Ecosystem Dynamics: Wormspace, Chaos and the Implications for Ecological Risk Assessment", USEPA Regional Risk Assessment Annual Meeting, May 4, 1993, Atlanta, GA.

Since September 1992, we have also prepared and submitted seven manuscripts, three of which are now in press. We have also sent out over 50 copies of these papers to various people interested in this research. Copies of these papers are presented in Appendix A.

In year two the specific accomplishments met included:

- Completing SAM experiments using Jet-A, JP-4 and the initial data collection for the JP-8 experiment.

- Completing MFC microcosm experiments using the standard protocol for the toxicants Jet-A and JP-4.

- An extensive investigation into the degradation of the WSF materials in the SAM and MFC systems has led to the preliminary conclusion that the biological communities may release these materials into the media during decomposition, redosing the system.

- Completing two sets of MFC experiments modified to explore specific questions as to the design of multispecies toxicity tests.

- Derivation of a novel method to examine ecological dynamics at the community and ecosystem level, the space time worms.

- Incorporation of nonlinear dynamics and chaos into the interpretation of ecosystem dynamics due to anthropogenic inputs.

- Improvements to the RIFFLE program, providing a graphical user interface so that nonmetric clustering and its association analysis can be accomplished without extensive programming.

- Application of these results to ecological risk assessment, including the conclusion that risk assessments are more akin to weather forecasts, that is forecasts with specified time limits that deal with a chaotic system.

Below is a more detailed summary of our research program from June 1, 1992 to May 31, 1993.

Overview of the Methodology

Toxicants. Jet-A, JP-4 and JP-8 are the toxicants for these studies. The Jet-A has been obtained from a commercial supplier, Chevron. The military fuels have been obtained from the U.S. Air Force Laboratories at Wright-Patterson AFB and are labeled as to lot number. Records and archival samples are maintained by the Quality Assurance program of the Institute.

Microcosm Protocol. The 64 day SAM protocol as developed by Taub (Taub *et al.*, 1988) consists of ten algal, four invertebrate and one bacterial species introduced into 3 L of sterile defined medium. Test containers are 4 L glass jars. An autoclaved sediment consisting of 200 g silica sand and 0.5 g of ground chitin are added to the already autoclaved jar and media. All complex toxicants are tested by removing 450 ml of media and organisms at the end of the 7 day acclimation period and adding appropriate amounts of jet fuel WSF and clean media that results in the final concentrations of toxicant. Concentrations for the tests run to date are 0, 1, 5 and 15 percent WSF. Numbers of organisms, dissolved oxygen (DO) and pH are determined twice weekly. Nutrients (nitrate, nitrite, ammonia, and phosphate) are sampled and measured twice weekly for the first four weeks, then only once weekly thereafter. A summary of the SAM methodology is presented in Table 1.

Mixed Flask Culture. The MFC microcosms are smaller systems of approximately 1 L and are inoculated with 50 ml of a stock culture originally derived from a natural system. The inoculum will be derived from the pond that is on the property of the Shannon Point Marine Center of WWU. Sand is also added to enhance the benthic populations included in the inoculum. Other variables to be measured include pH, DO so that a P/R ratio can be obtained, algae, total zooplankton, and ciliate protozoa.

Modifications to the original protocol have been made as part of additional studies conducted by R. Sandberg and S. Rodgers. In a study determining the applicability of the MFC when used to examine sediment contamination, R. Sandberg dosed the MFC by injecting jet fuel into the sediment. S. Rodgers is attempting to determine the importance of system complexity and similarity in the reproduction of results in the MFC system. In one set of experiments, only the SAM organisms were added by the normal cross inoculation to attempt to ensure homogeneity between replicates was performed. In a second set of experiments the SAM organisms were used but no cross inoculation. Summaries of these experiments are presented below. A summary of the NMC methodology is presented in Table 2.

Table 1. Summary of Test Conditions for a Typical Standardized Aquatic Microcosm
ASTM E 1366 - 91

ORGANISMS	
Type and number of test organisms per chamber:	Algae (added on Day 0 at initial concentration of 10^3 cells for each algae species): *Anabaena cylindrica, Ankistrodesmus* sp., Chlamydomonas *reinhardi* 90, *Chlorella vulgaris, Lyngbya* sp. *Nitzschia kutzigiana* (Diatom 216), *Scenedesmus obliquus, Selenastrum capricornutum, Stigeoclonium* sp., and *Ulothrix* sp. Animals (added on Day 4 at the initial numbers indicated in parentheses): *Daphnia magna* (16/microcosm), *Hyalella azteca* (12/microcosm), *Cypridopsis* sp. or *Cyprinotus* sp. (ostracod) (6/microcosm), Hypotrichs [protozoa] (0.1/mL) (optional), and *Philodina* sp. (rotifer) (0.03/mL)
EXPERIMENTAL DESIGN	
Test vessel type and size:	One-gallon (3.8 L) glass jars are recommended; soft glass is satisfactory if new containers are used; measurements should be 16.0 cm wide at the shoulder, 25 cm tall with 10.6 cm openings
Medium volume:	500 mL added to each container
Number of replicates x concentrations	6x4
Reinoculation:	Once per week add one drop (circa 0.05 mL) to each microcosm from a mix of the ten species = 5×10^2 cells of each alga added per microcosm
Addition of test materials:	Add material on Day 7; test material may be added biweekly or weekly after sampling
Sampling frequency:	2 times each week until end of test
PHYSICAL AND CHEMICAL PARAMETERS	
Temperature:	Incubator or temperature controlled room is required providing an environment 20 to 25°C with minimal dimensions of 2.6 by 0.85 by 0.8 m high
Light intensity:	80 µE m^{-2} photosynthetically active radiation s^{-1} (850 to 1000 fc)
Photoperiod:	12 h light / 12 h dark
Microcosm medium:	Medium T82MV adjusted to pH 7
Sediment:	Composed of silica sand (200 g), ground, crude chitin (0.5), and cellulose powder (0.5 g) added to each container
Typical Endpoints:	Population dynamics of each species, chemical-physical parameters, nutrients, diversity, predator-prey interactions, chemical fate.

Table 2. Summary of Test Conditions for Mixed Flask Culture Microcosms

TEST TYPE	Multispecies
ORGANISMS	
Number and type of organism:	a) two species of single-celled green algae or diatoms b) one species of filamentous green alga c) one species of nitrogen - fixing blue - green alga d) one grazing macroinvertebrate e) one benthic, detritus - feeding macroinvertebrate f) bacteria and protozoa species
EXPERIMENTAL DESIGN	
Test vessel type and size: Volume/Mass:	1 L beakers covered with a large petri dish 50 mL of acid washed sand sediment and 900 mL of Taub # 82 medium [20], into which 50 mL of inoculum was introduced
Number of groups:	4
Number of replicate chambers per group:	5
Reinoculation:	10 mL of stock community each week
Test duration:	12 - 18 weeks Allow to mature 6 weeks prior to treatment; follow 6 to 12 weeks after exposure
PHYSICAL AND CHEMICAL PARAMETERS	
Temperature:	20°C
Photoperiod:	12 h light / 12 h dark
Endpoint:	Oxygen content, algal densities, microbial activity, respiratory activity, biomass, protozoan populations

Sampling and Data Collection Procedures. All microcosm data are recorded onto a Macintosh Classic, hard copy printed, checked for accuracy and archived. The information is then fed into the Macintosh compatible data analysis system. Parameters calculated included the DO, DO gain and loss, nutrient concentrations, net photosynthesis/respiration ratio (P/R), pH, algal species diversity, daphnid fecundity, algal biovolume and biovolume of available algae. The statistical significance of each of these parameters compared to the controls are computed for each sampling day using the methodology of Conquest and Taub (1989).

Gas Chromatography of WSF. This protocol utilizes a Tekmar LSC 2000 Purge and Trap (P&T) concentrator system in tandem with a Hewlett Packard 5890A Gas Chromatograph with a Flame Ionization Detector (FID) (ASTM D3710, 1988; ASTM D2887, 1988; Westendorf, 1986). Instrument blanks and deionized distilled water blanks are used to verify the P&T and GC columns cleanliness prior to analysis of samples. A five mL sample is injected into a five milliliter sparger, purged with pre-purified nitrogen gas for eleven minutes and dry purged for four minutes. Volatile hydrocarbons, purged from the sample and collected on the Tenax/Silica Gel column, are desorbed at 180°C directly onto the gas chromatograph SPB-5, 30m x 0.53 mm ID 1.5µm film, fused silica capillary column. The column, at 35°C, is held at that temperature for two minutes, increased to 225°C at 12°C/min and held at that temperature for five minutes. A Spectra-Physics 4290 Integrator records the FID signal output of the volatile hydrocarbons that have been separated and eluted from the column by molecular weight.

Identification and quantification of GC fractions. Qualitative identification of some components in the water soluble fraction (WSF) of the JP-4 fuel, used as the toxicant in the microcosm test, were determined using a Simulated Distillation (SIMDIS) Calibration Mixture. The ASTM Method D3710 Qualitative Calibration Mixture is the standard test method for determining the Boiling Range Distribution of Gasoline and Gasoline Fractions by Gas Chromatography. This mixture was used as a calibration standard to determine the retention times for each known component in the mixture against which unknown components, in the WSF of the Jet fuel mixture, were compared and identified.

Quantitative estimates of some components of the WSF were made by comparing sample chromatographs to certified n-paraffin and n-naphtha chromatograph standards, prepared and analyzed under the same P&T/GC conditions.

Multivariate Techniques-Nonmetric Clustering. In the research described above, three multivariate significance tests were used. Two of them were based on the ratio of multivariate metric distances within treatment groups vs. between treatment groups. One of these is calculated using Euclidean distance and the other with cosine of vectors distance (Good, 1982; Smith *et al.*, 1990). The third test used nonmetric clustering and association analysis (Matthews and Matthews, 1990). In the microcosm tests there were four treatment groups with six replicates, giving a total of 24. This example is used to illustrate the applications in the derivations that follow.

Treating a sample on a given day as a vector of values, $\bar{\mathbf{x}} = \langle \mathbf{x}_1, \ldots \mathbf{x}_{17} \rangle$, with one value for each of the measured biotic parameters, allows multivariate distance functions to be computed. Euclidean distance between two sample points $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ is computed as

$$\sqrt{\sum_i (x_i - y_i)^2}$$

The cosine of the vector distance between the points $\bar{x}$ and $\bar{y}$ is computed as

$$1-\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum y_i^2}}$$

Subtracting the cosine from one yields a distance measure, rather than a similarity measure, with the measure increasing as the points get farther from each other.

The within-between ratio test used a complete matrix of point-to-point distance (either Euclidean or cosine) values. For each sampling date, one sample point $\bar{x}$ was obtained from each of six replicates in the four treatment groups, giving a 24 x 24 matrix of distances. After the distances were computed, the ratio of the average within group metric (*W*) to the average between group metric (*B*) was computed (*W/B*). If the points in a given treatment group are closer to each other, on average, than they are to points in a different treatment group, then this ratio will be small. The significance of the ratio is estimated with an approximate randomization test (Noreen, 1989). This test is based on the fact that, under the null hypothesis, assignment of points to treatment groups is random, the treatment having no effect. The test, accordingly, randomly assigns each of the replicate points to groups, and recomputes the *W/B* ratio, a large number of times (500 in our tests). If the null hypothesis is false, this randomly derived ratio will (probably) be larger than the *W/B* ratio obtained from the actual treatment groups. By taking a large number of random reassignments, a valid estimate of the probability under the null hypothesis is obtained as $(n+1)/(500+1)$, where n is the number of times a ratio less than or equal to the actual ratio was obtained (Noreen, 1989).

In the clustering association test, the data are first clustered independently of the treatment group, using nonmetric clustering and the computer program RIFFLE (Matthews and Hearne, 1991). Because the RIFFLE analysis is naive to treatment group, the clusters may, or may not correspond to treatment effects. To evaluate whether the clusters were related to treatment groups, whenever the clustering procedure produced four clusters for the sample points, the association between clusters and treatment groups was measured in a 4 x 4 contingency table, each point in treatment group i and cluster j being counted as a point in frequency cell ij. Significance of the association in the table was then measured with Pearson's X^2 test, defined as

$$\chi^2 = \sum_{ij} \frac{\left(N_{ij} - n_{ij}\right)^2}{n_{ij}}$$

where N_{ij} is the actual cell count and n_{ij} is the expected cell frequency, obtained from the row and column marginal totals N_{+j} and N_{i+} as

$$n_{ij} = \frac{N_{+j}N_{i+}}{N}$$

where $N = 24$ is the total cell count (Press *et al.*, 1990), and a standard procedure for computing the significance (probability) of χ^2, taken from Press *et al.* (1990).

Summary of Results to May 31, 1993

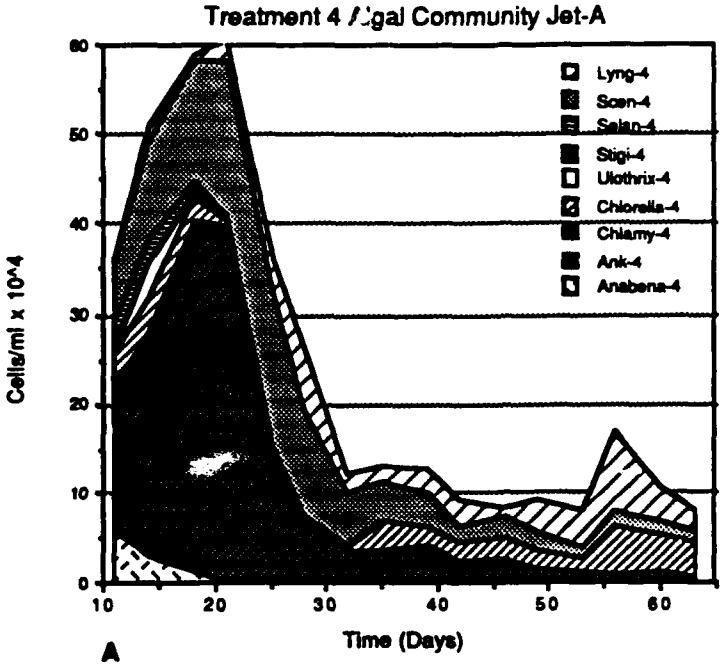
Summary of the Jet-A and JP-4 SAM experiments

Persistence of the fuels. In the case of both WSFs, within three weeks after dosing the original material had been volitilized or degraded. In the case of JP-4, benzene, 2,4 dimethylpentane, ethylbenzene, 2-methylpentane, 2-methylpropane, o-xylene and toluene, were tracked using GC analysis during the course of the SAM experiment. After week three, only 2-methylpentane and 2-methylpropane are detectable. Since only the 2-methylpropane is present 672 hours after dosing, this material may be the final biodegradative product of the absorbed fraction of the WSF, and is being investigated in more detail.

Comparison of Algal Population Dynamics-Highest Treatment. These area graphs (Fig. 1) show the contribution of each algal species to the algal assemblage for the highest treatment concentration for each experiment. In the Jet-A treatment the algal populations were highest, reflecting the increased toxicity of the Jet-A to the daphnid populations. In both experiments however, an algal bloom was observed during the first 30 days of the experiment. At the end of the experiment the numbers and composition of the algal assemblage were similar, although the proportions of the species making up the assemblage had some differences. Chlorella seemed to be a greater constituent of the community in the JP-4 experiment.

Daphnid Population Dynamics. The most direct effect of the jet fuel upon the population dynamics of the daphnid populations was the delay in daphnid reproduction (Fig. 2). Peaks were delayed in the Treatment 4 microcosms in both instances. Daphnids were very important in determining the clusters in the early part of each experiment but not as important later. In both experiments two peaks of daphnid populations are observed. The first reflects the presence of the toxicant, the second occurs similarly in the dosed and not dosed systems. Error bars are not shown for clarity.

Ostracod Population Dynamics. Ostracod populations did not increase until late in each experiment (Fig. 3). In the Jet-A experiment (A), the numbers started an increase between days 40 and 45.



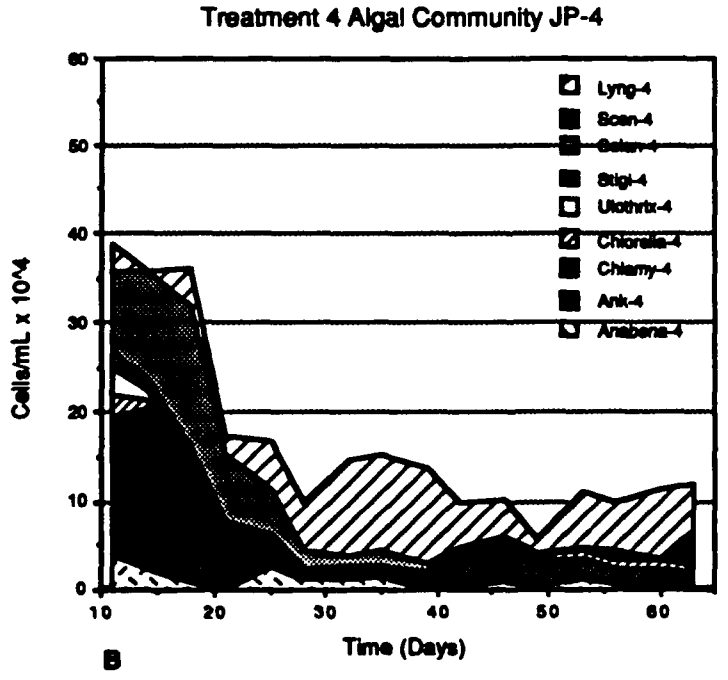
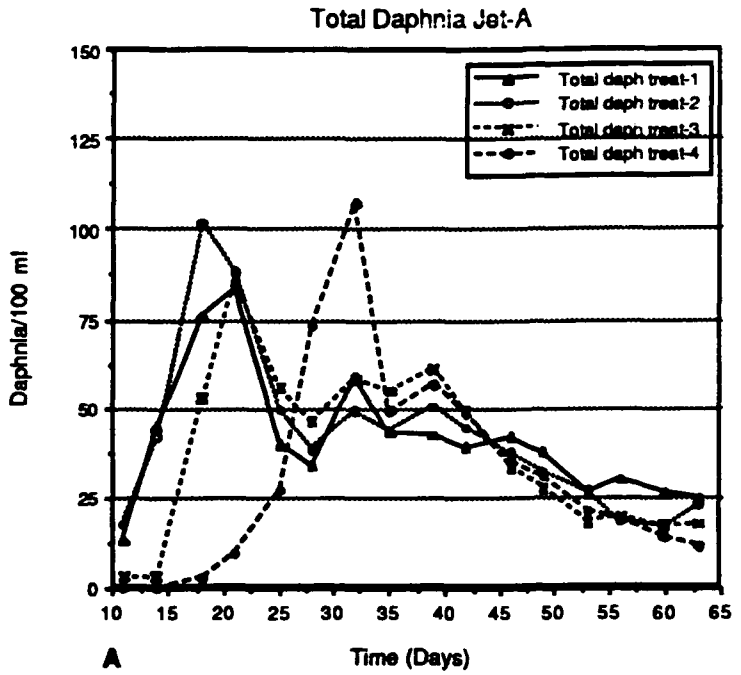


FIG. 1--Comparison of algal population dynamics-highest treatment.



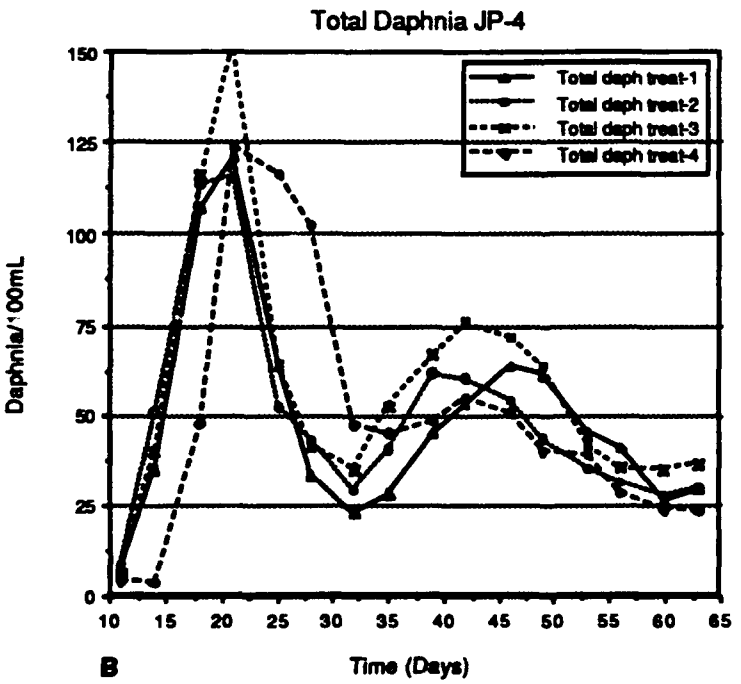


FIG. 2--Daphnid population dynamics.

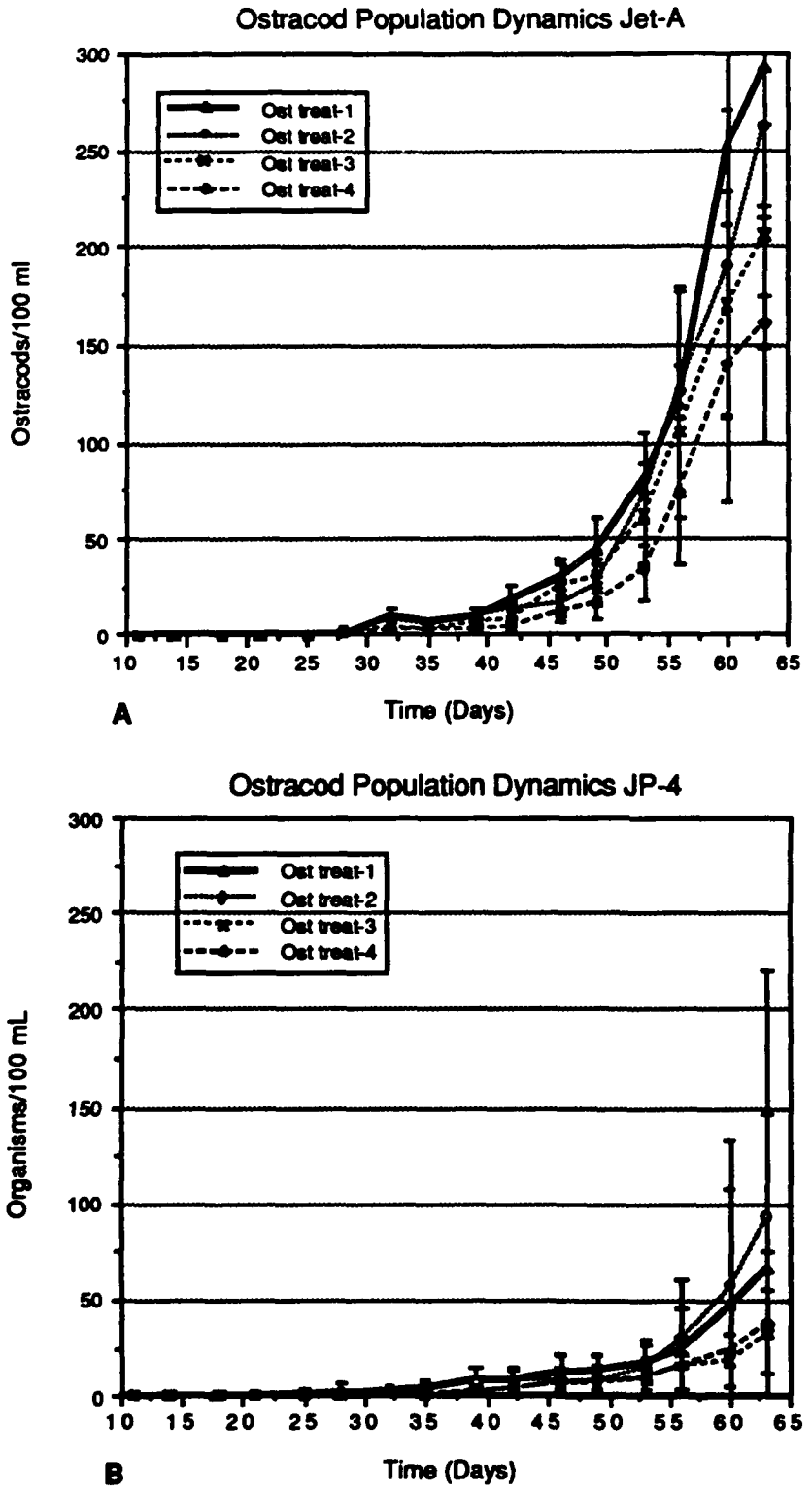


FIG. 3--Ostracod population dynamics.

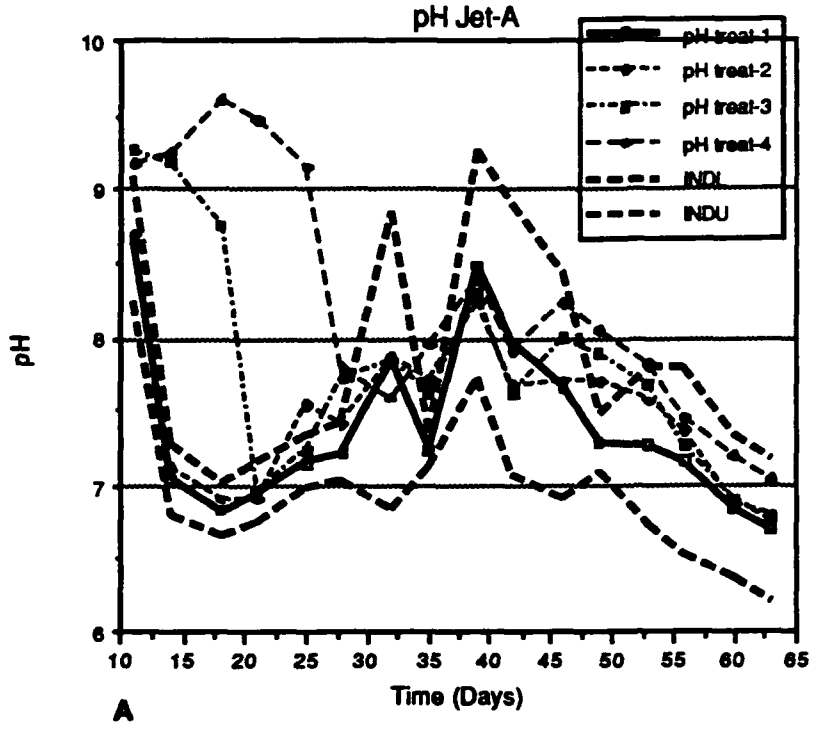
The experiment using JP-4 as a toxicant (B) did not see the increase in ostracods until between days 50-55, approximately ten days later. Consequently, the total numbers of ostracods observed were not as high in the JP-4 microcosms. Note that the order of densities in the Jet-A experiment followed a dose response pattern, as did the JP-4 experiment, even with the lower total numbers. Conventional analysis did not demonstrate significance, however nonmetric clustering did indicate the importance of the ostracods in determining clusters in both sets of microcosm experiments.

Philodina Population Dynamics. Philodina did not become prevalent in the microcosms until the second half of the experiment. One of the major problems was the inherent variability in the sampling and in the replicates. Organisms that reproduce rapidly can show large differences in population sizes during the course of a sampling day. Although, in the later stages of the microcosm experiments the dosed systems had a generally larger number of the rotifers, the results were not statistically significant using conventional IND plots. However, using cluster analysis, Philodina were also determined to be an important variable in defining clusters. This held true for both the Jet-A and JP-4 experiments.

Comparisons of pH dynamics of the Jet-A and JP-4 Experiments. Unlike the biotic variables, pH did reflect some of the oscillations detected by the cluster analysis (Fig. 4). In both the Jet-A and the JP-4 experiments the highest concentrations demonstrated a statistically significant difference, determined by the interval of non-significant difference during the first 30 days of the experiment. The second oscillation, between days 45 and 50, is not as clear since only one sampling date demonstrated the statistically significant difference. Type II error becomes a concern with so many comparisons, even with the corrections incorporated into the IND plots.

Photosynthesis/Respiration Ratio. The photosynthesis/respiration ratio reflects the oscillations seen in pH and the clustering analysis for the first 30 days and then only for the Jet-A water soluble fraction. In the Jet-A experiment, a second deviation from the IND plot was noted in the period corresponding to the second oscillation, but the result is difficult to distinguish from a type II error. In the JP-4 experiment, the IND plots are large, reflecting the variance in those sampling days. As an "emergent property", it is not clear if the P/R ratio provides any more information in this experiment than the clustering based upon the biotic components.

Oscillations in Community Dynamics Observed in both the Jet-A and the JP-4 Experiments. The Jet-A and the JP-4 SAM experiments both displayed a series of oscillations; revealed by the three clustering techniques employed in the analysis (Fig. 5). The first oscillation, as defined by Cosine Distance common to each experiment, is due to the interaction of the daphnid population and the algae. The result is statistically significant, as determined by the goodness-of-fit confidence level, graphed by day in Fig. 6. In both experiments, the oscillation is within the first 30 days of the SAM time-line. Interestingly, the magnitude of the first oscillation, as determined by Cosine Distance, is less in the JP-4 experiment, possibly reflecting the reduced acute and chronic toxicity of the mixture.



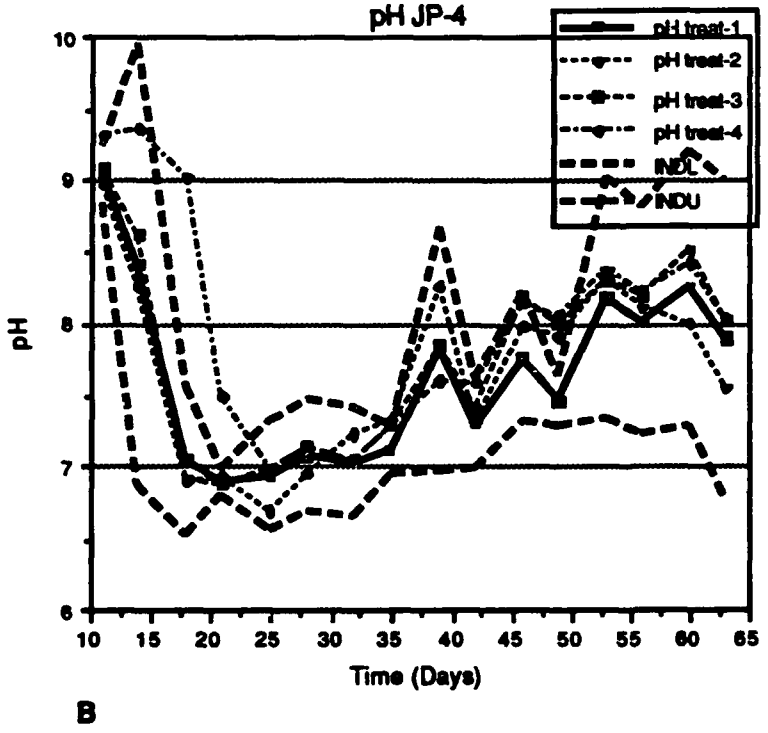


FIG. 4--Comparisons of pH during the SAM studies.

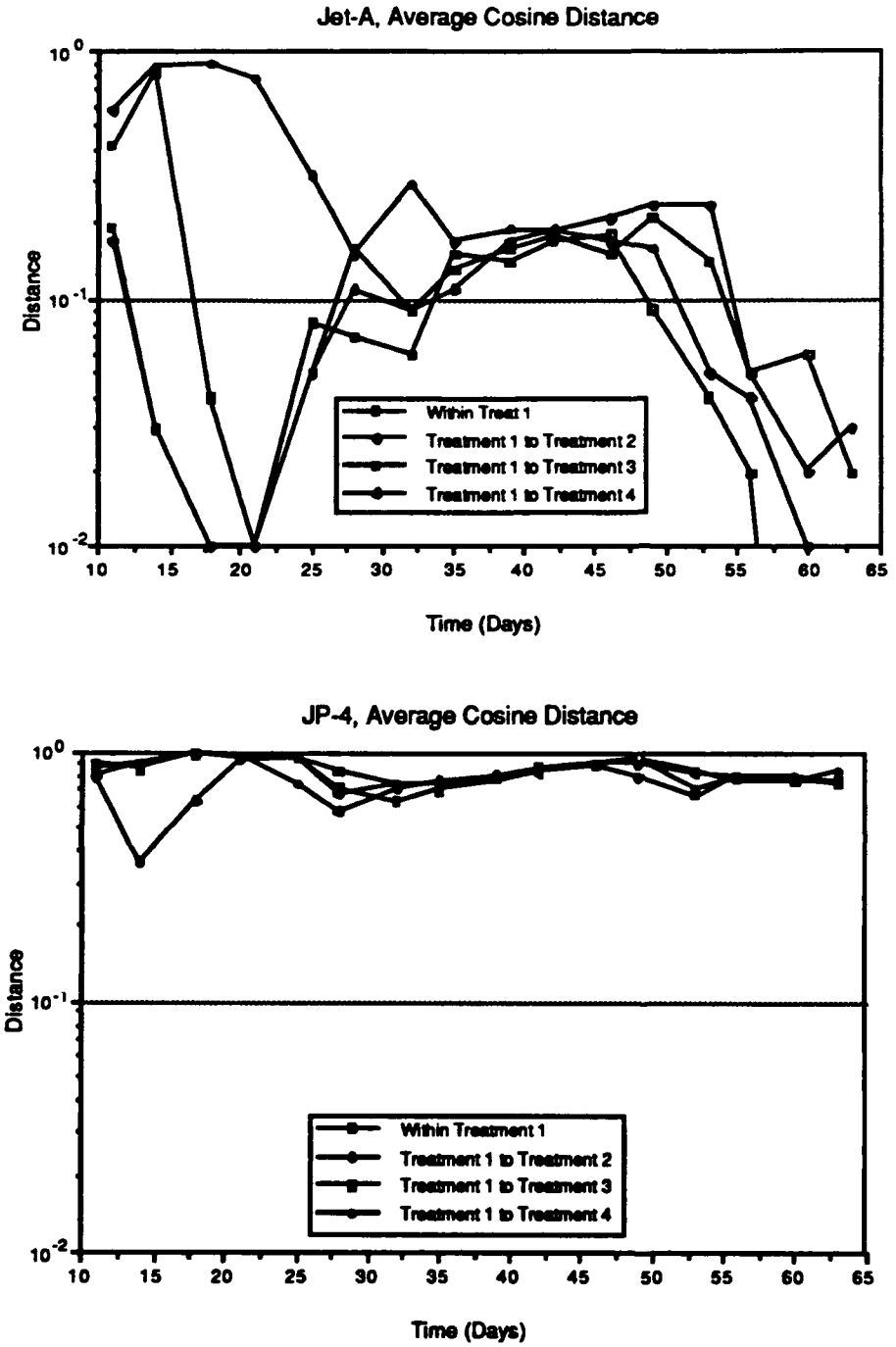
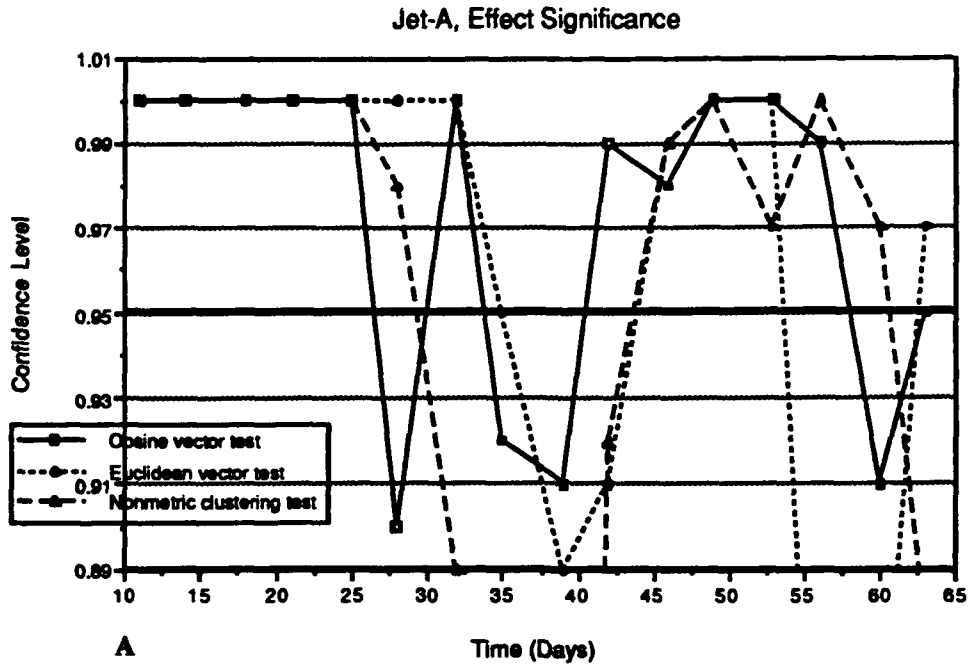


FIG. 5--Cosine Distances of the Treatment 1 to the dosed treatments in the Jet-A and the JP-4 SAMs.



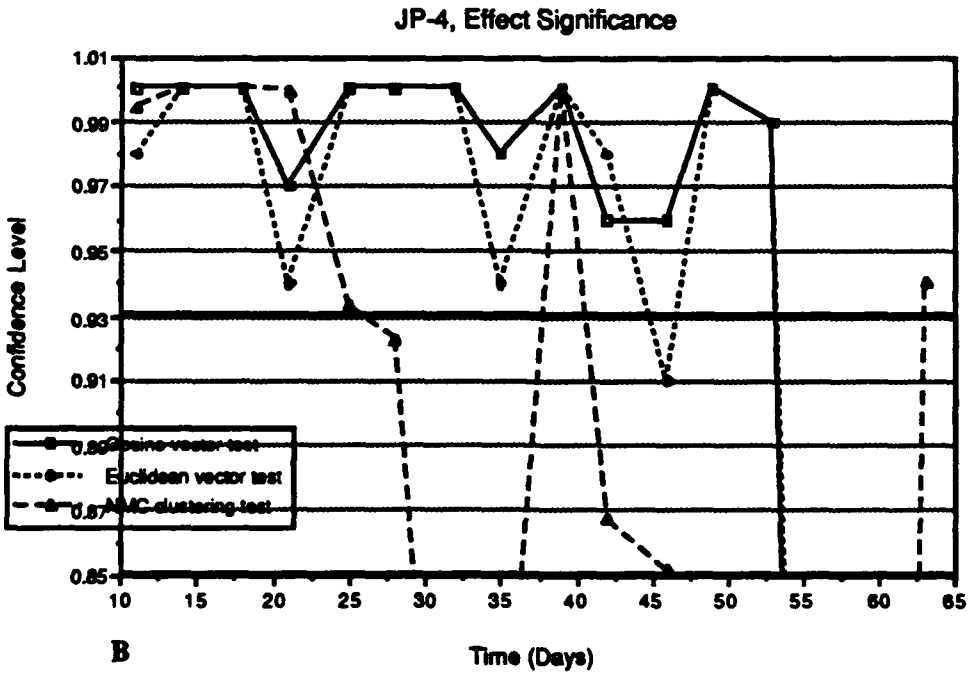


FIG. 6--Significance of the association analysis of the 4 Treatments in the Jet-A and the JP-4 SAMs.

A second series of oscillations, as measured by Cosine Distance, occur in the last thirty days of each experiment. Again the oscillations are statistically significant.

The participants in the community that contribute to these oscillations are slightly different judging by the table of important variables (Table 3). Unfortunately, the length of the SAM protocol is not sufficient to conduct an analysis of the period and amplitude of the oscillations. Another complication in examining the results is the difficulty in making direct comparisons between experiments. Although the Cosine Distance may be the same, the orientation of the angle can be quite different.

Table 3. Variable ranking by success in determining clusters as defined by nonmetric clustering. Variables such as Ankistrodesmus and the Daphnia classes ranked highly in the course of this study. However, reliance on any particular organism or a small combination of variables would inadequately describe the dynamics of the system.

Jet-A		JP-4	
Variable	Ranked	Variable	Ranked
Ankistrodesmus	12	Chlorella	8
M. Daphnia	11	S. Daphnia	8
Chlorella	9	Ankistrodesmus	6
Scenedesmus	7	Scenedesmus	5
S. Daphnia	6	Philodina	5
L. Daphnia	5	M. Daphnia	4
Ostracod	4	Lyngbya	4
Philodina	4	L. Daphnia	3
Selenastrum	4	Ostracod	3
Lyngbya	3	Selenastrum	3
Ulothrix	1		

Discussion

First, the apparent recovery or movement of the dosed systems towards the reference or treatment 1 case may be an artifact of our measurement systems that allow the n-dimensional data to be represented in a two dimensional system. In an n-dimensional sense, the systems may be moving in opposite directions and simply pass by similar coordinates during certain time intervals. Positions may be similar but the n-dimensional vectors describing the movements of the systems can be very different. A representation of these dynamics is presented in Fig. 7. The two systems intersect, although the vectors are quite different.

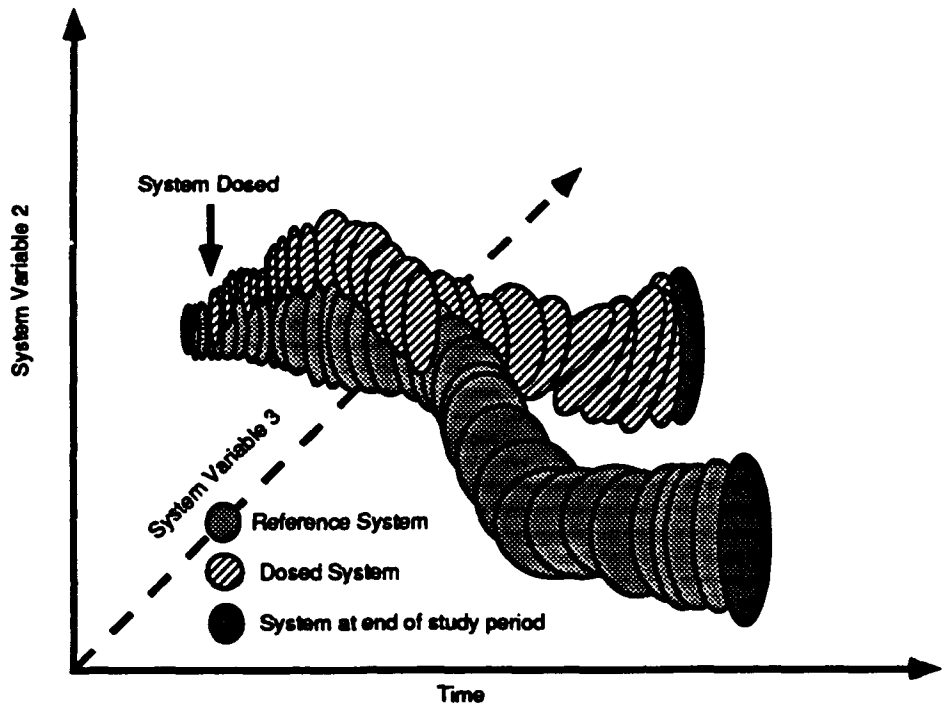


FIG. 7--Visualization of ecosystem dynamics to reflect a possible interpretation of the impacts of the jet fuels.

The apparent recoveries and divergences may also be artifacts of our attempt to choose the best means of collapsing and representing n-dimensional data into a two or three dimensional representation. In order to represent such data it is necessary to project n-dimensional data into three or less dimensions. As information is lost as the shadow from a cube is projected upon a two dimensional screen, a similar loss of information can occur in our attempt to represent n-dimensional data. Not every divergence from the reference treatment may have a cause directly related to it in time. Differentiating those events from those due to degradation products or other perturbations is challenging.

Not only may system recovery be an illusion, but there are strong theoretical reasons that seem to indicate that recovery to a reference system may be impossible or at least unlikely. In fact, systems that differ only marginally in their initial conditions and at levels probably impossible to measure are likely to diverge in unpredictable manners. May and Oster (1978) in a particularly seminal paper investigated the likelihood that many of the dynamics seen in ecosystems that are generally attributed as chance or stochastic events are in fact deterministic. In fact, simple deterministic models of populations can give rise to complex dynamics. Using equations resembling those used in population biology, bifurcations occur resulting in several distinct outcomes. Eventually, given the proper parameters, the system

appears chaotic in nature although the underlying mechanisms are completely deterministic. Obviously, biological systems have limits, extinction being perhaps the most obvious and best recorded. Another ramification is that the noise in ecosystems and in sampling may not be the result of a stochastic process but the result of underlying deterministic, but chaotic relationships.

These principals also apply to spatial distributions of populations as recently reported by Hassell *et al.* (1991). In a study using host-parasite interactions, a variety of spatial patterns were developed using the Nicholson-Bailey model. Host-parasite interactions demonstrated dynamics ranging from static 'crystal lattice' patterns, spiral waves, chaotic variation, or extinction with the appropriate alteration of only three parameters within the same set of equations. The deterministically determined patterns could be extremely complex and not distinguishable from stochastic environmental changes.

Given the perhaps chaotic nature of populations it may not be possible to predict species presence, population interactions, or structural and functional attributes. Katz *et al.* (1987) examined the spatial and temporal variability in zooplankton data from a series of five lakes in North America. Much of the analysis was based on limnological data collected by Brige and Juday from 1925 to 1942. Copepods and cladocera, except *Bosmina*, exhibited larger variability between lakes than between years in the same lake. Some taxa showed consistent patterns among the study lakes. They concluded that the controlling factors for these taxa operated uniformly in each of the study sites. However, in regards to the depth of maximal abundance for calanoid copepods and Bosmina, the data obtained from one lake had little predictive power for application to other lakes. Part of this uncertainty was attributed to the intrinsic rate of increase of the invertebrates with the variability increasing with a corresponding increase in r_{max}. A high r_{max} should enable the populations to accurately track changes in the environment. Katz et al suggest that these taxa be used to track changes in the environment. Unfortunately, in the context of environmental toxicology, the inability to use one "reference" lake to predict the non-dosed population dynamics of these organisms in another eliminates comparisons of the two systems as measures of anthropogenic impacts.

A better strategy may be to let the data and a clustering protocol identify the important parameters in determining the dynamics of and impacts to ecological systems. This approach has been recently suggested independently by Dickson *et al.* (1992), Matthews *et al.* 1991, and Matthews and Matthews 1991. This approach is in direct contrast to the more usual means of assessing anthropogenic impacts. One classical approach is to use the presence or absence of so called indicator species. This assumes that the tolerance to a variety of toxicants is known and that chaotic or stochastic influences are minimized. A second approach is to use hypothesis testing to differentiate metrics from the systems in question. This second approach assumes that the investigators know *a priori* the important parameters to measure. Given that in our relatively simple SAM systems that the important parameters in differentiating non-dosed from dosed systems change from sampling period to sampling period, this assumption can not

be made. Classification approaches such as nonmetric clustering or the canonical correlation methodology developed by Dickson *et al.* (1992), eliminates these assumptions.

These results presented in this report and by others reviewed above and the implications of chaotic dynamics suggest that reliance upon any one variable or an index of variables may be an operational convenience that may provide a misleading representation of pollutant effects and associated risks. The use of indices such as diversity and the Index of Biological Integrity have the effect of collapsing the dimensions of the descriptive hypervolume. Indices, since they are composited variables, are not true endpoints. The collapse of the dimensions that are composited tends to eliminate crucial information, such as the variability in the importance of variables. The mere presence or absence and the frequency of these events can be analyzed using techniques such as nonmetric clustering that preserve the nature of the dataset. A useful function was certainly served by the application of indices, but the new methods of data compilation, analysis and representation derived from the Artificial Intelligence tradition can now replace these approaches and illuminate the underlying structure and dynamic nature of ecological systems.

The implications are important. Currently, only small sections of ecosystems are monitored or a heavy reliance is placed upon so called indicator species. These data suggest that to do so is dangerous, may produce misleading interpretations resulting in costly error in management and regulatory judgments. Much larger toxicological test systems are currently analyzed using conventional statistical methods on the limit of acceptable statistical power. Interpretation of the results has proven to be difficult, if not confusing. Application of the approach and tools that proved successful in revealing the complex dynamics of these small microcosms should prove useful in analyzing larger toxicological test systems and field research.

CONCLUSIONS

(1) In both of the experiments, multiple oscillations of the dosed treatment groups away from the reference treatment were observed using multivariate statistics. The first oscillation is due to the differential impact of the WSF of the jet fuels to the algae-daphnid population dynamics. The following oscillations, although statistically significant and seen in both experiments, is not as clear cut. The divergence of the second oscillation may be due to two separate mechanisms.

(a) A fluctuation due to the initial stress has occurred, but in such a fashion that an incompletely dampened oscillation repeats. There has been no fundamental alteration in the functioning of the ecosystem, and the oscillations are a result of the inherent time lags and stochastic factors governing the dynamics of the system.

(b) A fundamental aspect of the ecosystem has been altered so that the repeated oscillations reflect the persistence of the impact. An alteration in the detritus quality or in the community involved in the recycling of detritus may have long term impacts as other nutrients become limiting in the system. Nutrients are at low levels during the second 30 days of a typical SAM experiment. This possibility could include a fundamental and long lasting effect upon the system, contrary to the first mechanism.

(2) A combination of multivariate analyses appear to be useful and illuminating in assessing the long term dynamics of these systems. Each has strengths that make multivariate analysis a strong methodology with powerful advantages to conventional univariate methods.

(3) Although simple systems, the SAM experiments exhibits complex dynamics and behaviors. The protocol results in a persistent system with good replicability within an experiment, even with complex species interactions.

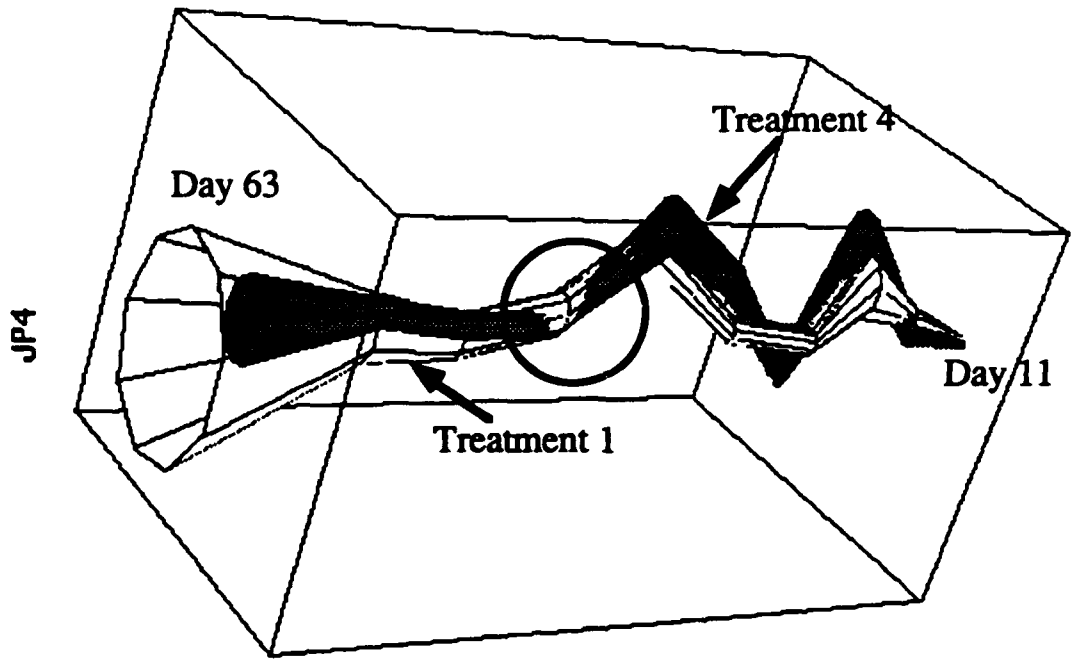
(4) Techniques that allow the reduction and visualization of even these relatively simple multispecies toxicity tests should contribute to our understanding of system dynamics and improve hazard assessment.

Research in Progress-Summaries

Comparison of Test Results in the Evaluation of the WSF of Several Jet Fuels Using the Standardized Aquatic Microcosm and the Mixed Flask Culture Protocols. The water soluble fraction of the turbine fuels Jet-A, JP-4 and JP-8 have been examined as stressors for two microcosm protocols, the standardized aquatic microcosm (SAM) and the mixed flask culture (MFC). The SAM is a 3 L system inoculated with standard cultures of algae, zooplankton, bacteria, and protozoa. In contrast, the MFC is 1 L and is inoculated with a complex mixture of organisms derived from a natural source. Analysis of the organism counts and physical data were conducted using conventional and newly derived multivariate methods. Physical parameters, such as pH and oxygen metabolism, were often not as sensitive as species and bacterial counts. Like the SAM system, species numbers and other variables that determined clusters varied among sampling dates. Compared to the larger yet simpler system, the MFC exhibits more violent dynamics and is more likely to become catastrophically fixated, as in systems dominated by cyanobacteria. The combination of greater diversity and smaller volume may contribute to the volatile or chaotic dynamics of the MFC system.

Response Volumes (Space-time Worms) as a Method for the Visualization of Ecosystem Dynamics and Indirect Effects. A variety of indexes and other composite measures of ecosystems, such as measures of integrity and diversity, have been used to summarize the state of an ecosystem. These approaches have numerous shortcomings. We have developed a method for the visualization and

quantification of the state of an ecosystem that projects from the original n-dimensional space into a two dimensional representation. Currently, a principal components projection provides the axes to plot the system in a two dimensional space. In studies with several sampling dates, a projection is plotted for each sampling day and then connected to form a three dimensional representation of the changes of the ecosystem over time (Fig. 8). The response-volumes or space-time worms generated by this process provide a three dimensional representation of the changes of an ecosystem over time. Various perspectives can be generated until the best viewing point is selected for the particular attribute or question under consideration. The method has proven vital in the examination of microcosm ecosystems dosed with a variety of toxicants and should prove useful in the analysis of FIFRA type microcosms and various field studies.



Response Area (Wormspace) for the JP-4 SAM Experiment

FIG. 8--Space-time worms for the non-dosed (treatment 1) and highest dosed (treatment 4) systems of 6 replicates.

Non-linear Dynamics of Microcosm Ecosystems and the Inherent Limitations of Risk Assessment. Projections into two dimensional space with time are used to visualize ecosystem dynamics. The space-time worm projections have demonstrated that the systems are moving in a complex dynamic that does not repeat or recover as defined as the return of the dosed system to the space and dynamics of the non-dosed case. In cases where the dosed and non-dosed treatments overlap, the subsequent dynamics demonstrated that it is a case of passing through and not recovery. The patterns appear to be chaotic,

such as turbulence and weather. Ecological important properties of these systems are: they do not return to an original condition upon perturbation; the history of the perturbation resets the initial conditions making a return to the initial state virtually impossible; history of the system is important in setting the potential dynamics; and that predictions are limited not by knowledge but by the inherent dynamics of the system. Risk assessments and projections of impacts upon populations and communities have inherent limits on their power of prediction. These limits are inherent to the underlying dynamics of the system and not based on the uncertainty of the available knowledge.

Characterization and Classification of Direct and Indirect Effects at the Community and Ecosystem Levels. The dynamics of the response of an ecosystem to a stressor have classically been separated into direct and indirect effects. The initial direct effects of a toxicant alter the community in two ways. First, the system can be displaced from its initial state. The magnitude of the displacement may be estimated using current laboratory toxicity tests, however, given the complexity or even chaotic nature of ecosystems, the directional vector of this displacement may be impossible to predict. Second, the dispersion or variability of the system can also be altered. In some instances the variability of the system can be radically decreased or increased depending upon the type of toxicant. Indirect effects, however, may be so persistent as to take another stressor event to remove the impacts of this history from the system. In our studies, recovery in the classical sense of returning to the original or reference state is unlikely to occur. Even in unstressed systems small initial differences give rise to dramatic changes. The accurate prediction of direction and magnitude of the indirect effect may prove impossible if ecosystems exhibit sufficiently complex or chaotic dynamics.

Graduate Student Projects

Use of the Mixed Flask Culture (MFC) Microcosm Protocol to Investigate the Effects of a Pulsed Release of Jet-A--R.S. Sandberg and M.J. Roze. A 60-day 1 L Mixed Flask Culture (MFC) microcosm utilizing organisms derived from natural systems was used to assess the potential ecosystem level effects of a simulated release of a complex hydrocarbon mixture from sediments. A spiked layer of Standardized Aquatic Microcosm (SAM) sediment was encapsulated under an overlying layer of coadapted MFC silica sand and detritus. Treatment sediment groups consisting of six microcosm replicates were spiked with 0, 2, 10 and 25 microliters of Jet-A based on the results of preliminary acute 10-day freshwater sediment amphipod bioassays using *Hyalella azteca* as the test species. A slow, pulsed release of the test material from the spiked layer was obtained by stirring vigorously twice weekly throughout the test. Statistically significant effects among both community level physical properties and individual species population dynamics were observed using conventional univariate and multivariate techniques as well as a recently developed nonmetric multivariate clustering technique despite the relatively small proportion of Jet-A used in the test.

Evaluation of Community Structure and Community Function After Exposure to the Turbine Fuel Jet-A--S.C. Rodgers. The underlying premises of the Mixed Flask Culture (MFC), an aquatic microcosm design, include: 1) that the effects of a perturbation to an aquatic community may be monitored through the measurement of its functional parameters (i.e. pH and productivity/respiration ratio), and 2) these measurements will be similar between different wild-derived communities given the same perturbation. Two MFC experiments were conducted to assess these two premises. The treatment groups in both experiments consisted of 0%, 1%, 5%, and 15% WSF Jet-A with six replicates respectively. The experimental designs reflected both the MFC and the Standard Aquatic Microcosm (SAM); this hybrid design resulted in following a MFC protocol, but incorporated the SAM specified laboratory cultured organisms. Beaker heterogeneity was encouraged in the second experiment by not cross inoculating or reinoculating. The differences between the two experiments was designed to indicate if differently derived communities react similarly to an identical perturbation. Do the microcosms within each treatment group resemble each other functionally throughout the experiment, or is the within group deviation greater than the between group deviation?

Comparison of the Degradation of Water Soluble Components in Jet Fuel Using the Standard Aquatic Microcosm (SAM) and the Mixed Flask Microcosm (MFC).--A.J. Markiewicz. The Standard Aquatic Microcosm (SAM), a synthetic assemblage of organisms derived from laboratory cultures, was used in comparison with the Mixed Flask Microcosm (MFC), derived from natural sources, to monitor the degradation rates and biodegradation products of water soluble components in jet fuel and to evaluate whether ecosystem dynamics are similar between the two microcosm systems; independent of species diversity and trophic level complexity. The SAM microcosms were used for analysis of the water soluble fraction of JP-8, and the MFC microcosms were used for the water soluble fraction of Jet-A. Component degradation and by-products were monitored using Purge and Trap / Gas Chromatography. Preliminary results from both microcosms, using regression and multivariate analysis, indicate that all components are degraded simultaneously, but at different rates; component degradation rates oscillate in similar patterns temporally; most WSF components are completely degraded within 10-15 days; and that biodegradation products continue to reappear in a cyclic pattern throughout the experiment. In the SAM microcosms, WSF jet fuel components were rapidly sequestered from the water column and degradative rates were lower. Both microcosms form significantly distinct groups when clustered by degradation rates.

References and Bibliography

ASTM D3710 (1988) Standard Test Method for Boiling Range Distribution of Gasoline and Gasoline Fractions by Gas Chromatography, 1988 Annual Book of ASTM Standards, Vol. 5.03, pp 78-88. American Society for Testing and Materials, Philadelphia.

ASTM D2887 (1988) Standard Test Method for Boiling Range Distribution of Petroleum Fractions by Gas Chromatography, 1988 Annual book of ASTM Standards, Vol. 5.02, pp 506-513. American Society for Testing and Materials, Philadelphia.

ASTM E 1218 (1991) Conducting Static 96-h Toxicity Tests with Microalgae. Annual book of ASTM Standards, Vol. 11.04, pp 845-856. American Society for Testing and Materials, Philadelphia.

ASTM E 1366-91 (1991) Standard Practice for the standardized aquatic microcosm: fresh water, Vol 11.04. pp 1017-1051. American Society for Testing and Materials, Philadelphia.

Conquest, L.L. and Taub, F.B. (1989) Repeatability and reproducibility of the Standard Aquatic Microcosm: Statistical properties. In *Aquatic Toxicology and Hazard Assessment: 12th Volume, ASTM STP 1027* (Cowgill, U.M. and Williams, L.R., eds) American Society for Testing and Materials, Philadelphia, PA, pp. 159-177.

Crow , M.E. and Taub, F.B. (1979) Designing a microcosm bioassay to detect ecosystem level effects. *Intern. J. Environmental Studies.* 141-147.

Dickson, K.L., Waller, W.T., Kennedy, J.H. and Ammann, L.P. (1992) Assessing the relationship between ambient toxicity and instream biological response. *Env. Tox. Chem.* **11**, 1307-1322.

Fienberg, S.E. (1985) *The Analysis of Cross-Classified Categorical Data.* MIT Press, Cambridge, MA.

Fisher, L. (1992) Memorandum: Decisions on the Ecological, Fate and Effects Task Force. Office of Pesticides and Toxic Substances, U. S. Environmental Protection Agency.

Good, I.J. (1982) An index of separateness of clusters and a permutation test for its significance. *J. Statist. Comp. Simul.* **15**, 81-84.

Haley, M.V., Johnson, D.W. and Landis, W.G. (1988) The aquatic toxicity of brass dust. In *Aquatic Toxicology and Environmental Fate: Tenth Volume ASTM STP -971* (Adams, W., Chapman, G. and Landis, W.G., eds) American Society for Testing and Materials, Philadelphia. pp 468-479.

Hassell, M.P.H, Comins, N. and May, R.M. (1991) Spatial structure and chaos in insect population dynamics. *Nature* **353**, 255-258.

Johnson, A.R. (1988a) Evaluating ecosystem response to toxicant stress: a state space approach. In *Aquatic Toxicology and Hazard Assessment: 10th Volume, ASTM STP* 971 (Adams, W.J., Chapman, G.A. and Landis, W.G., eds) American Society for Testing and Materials, Philadelphia, pp. 275- 285.

Johnson, A.R. (1988b) Diagnostic variables as predictors of ecological risk. *Environmental Management* **12**, 515-523.

Katz, T.K., Frost, T.M. and Magnuson, J.J. (1987) Inferences from spatial and temporal variability in ecosystems: Long-term zooplankton data from lakes. *Amer. Nat.* **129**, 830-846.

Kersting, K. (1984) Development and use of an aquatic micro-ecosystem as a test system for toxic substances. Properties of an aquatic micro-ecosystem IV. *Int. Revue ges. Hydrobiol.* **69**, 567-607.

Kersting, K. (1985) Properties of an aquatic micro-ecosystem V. Ten years of observations of the prototype. *Verh. Internat. Verein Limnol.* **22**, 3040-3045.

Kersting, K. (1988) Normalized ecosystem strain in micro-ecosystems using different sets of state variables. *Verh. Internat. Verein. Limnol.* **23**, 1641-1646.

Kersting , K., and van Wungaarden, R. (1992) Effects of Chlorpyifos on a microecosystem. *Env. Tox, Chem.* **11**, 365-372.

Kindig, A.C., Loveday, L.C. and Taub, F.B. (1983) Differential sensitivity of new versus mature synthetic microcosms to streptomycin sulfate treatment. In *Aquatic Toxicology and Hazard Assessment: Sixth Symposium, ASTM 802.* (Bishop, W.E., Cardwell, R.D. and Heidolph, B.B., eds) American Society for Testing and Materials, Philadelphia, pp. 192-203.

Landis, W.G., Chester, N.A., Haley, M.V., Johnson, D.W., Muse, Jr., W.T. and Tauber, R.M. (1989) The utility of the standard aquatic microcosm as a standard method for ecotoxicological evaluation. In *Aquatic Toxicology and Environmental Fate: Eleventh Volume ASTM STP -1007* (Suter, G. and Adams, M., eds) American Society for Testing and Materials, Philadelphia pp 353-367.

Landis, W.G., Haley, M.V. and Chester, N.A. (1993) The use of the standardized aquatic microcosm in the evaluation of degradative bacteria in reducing impacts to aquatic ecosystems. In *Environmental Toxicology and Risk Assessment: First Volume, ASTM STP -1179* (Landis, W.G., Hughes, J. and Lewis, M., eds) *In press,* American Society for Testing and Materials, Philadelphia, in press.

Matthews, G.B. and Matthews, R.A. (1990) A model for describing community change. In *Pesticides in Natural Systems: How Can Their Effects Be Monitored? Proceeding of the Conference,* Environmental Research Laboratory/ORD, Corvallis, OR, EPA 9109/9-91/011.

Matthews, G.B. and Hearne, J. (1991) Clustering without a metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **13**, 175-184.

Matthews, G.B. and Matthews, R.A. (1991) A model for describing community change. In *Pesticides in Natural Systems: How Can Their Effects Be Monitored? Proceeding of the Conference,* Environmental Research Laboratory/ORD, Corvallis, OR, EPA 9109/9-91/011.

Matthews, G.B., Matthews, R.A. and Hachmoller, B. (1991) Mathematical analysis of temporal and spatial trends in the benthic macroinvertebrate communities of a small stream. *Canadian Journal of Fisheries and Aquatic Sciences.* **48**, 2184-2190.

Matthews, R.A., Matthews, G.B. and Ehinger, W.J. (1991) Classification and ordination of limnological data: a comparison of analytical tools. *Ecological Modeling.* **53**, 167-187.

May, R.M. and Oster, G.F. (1978) Bifurcations and dynamical complexity in simple ecological models. *Amer. Nat.* **110**, 573-599.

Noreen, E.W. (1989) *Computer Intensive Methods for Testing Hypotheses.* Wiley-Interscience, New York, NY.

Press, W.H., Flannery, B.P., Teukolsky, A.A. and Vetterline, W.T. (1990) *Numerical Recipes in C, the Art of Scientific Computing.* Cambridge University Press, New York, NY.

Smith, E.P., Pontasch, K.W. and Cairns, Jr., J. (1990) Community similarity and the analysis of multispecies environmental data: a unified statistical approach. *Water Res.* **24**, 507-514.

Sugiura, K. (1992) A multispecies laboratory microcosm for screening ecotoxicological impacts of chemicals. *Env. Tox. Chem.* **11**, 1217-1226.

Suter, G. (1993) A critique of ecosystem health: Concepts and indices. *Environ Tox. Chem.* in press.

Taub, F.B. (1969) Gnotobiotic models of freshwater communities. *Verh Internat. Verein. Limnol.* **17**, 485-496.

Taub , F.B. (1976) Demonstration of pollution effects in aquatic microcosms. *Intern J. Environmental Studies.* **10**, 23-33.

Taub, F.B. (1988) Standardized aquatic microcosm - development and testing. *Aquatic Ecotoxicology* II.

Taub, F.B. (1989) Standardized aquatic microcosms. *Environm. Sci. Technol.* 23, 1064-1066.

Taub, F.B. and Crow, M.E. (1978) Loss of a critical species in a model (laboratory) ecosystem. *Verh. Internat. Verein. Limnol.* 1270-1276.

Taub, F.B. , Crow, M.E. and Hartmann, H.J. (1980) Responses of aquatic microcosms to acute mortality. *Microcosms in Ecological Research*. Giesy, J.P. Jr., Technical Information Center, U. S. Department of Energy. Washington, D.C., 513-535.

Taub, F.B., Kindig, A.C. and Conquest, L.L. (1987) Interlaboratory testing of a standardized aquatic microcosm. In *Aquatic Toxicology and Hazard Assessment: 10th Volume, ASTM STP 971* (Adams, W.J., Chapman, G.A. and Landis, W.G., eds) American Society for Testing and Materials, Philadelphia, PA, pp. 385-405.

Taub, F.B., Kindig, A.C., Conquest, L.L. and Meador, J.P. (1988) Results of the interlaboratory testing of the Standardized Aquatic Microcosm protocol. In *Aquatic Toxicology and Hazard Assessment: Eleventh Symposium, ASTM* (Suter, G. and Lewis, M., eds) American Society for Testing and Materials, Philadelphia, PA.

Taub , F.B. and Read, P.L. (1983): Standardized Aquatic Microcosm Protocol: Final Report Contract No. 223-80-2352, Vol II. Food and Drug Administration. Washington, D.C.

Taub, F.B., Rose, K.A., Swartzman, G.L. and Taub, J.H. (submitted) Translating population toxicity to community effects. *Env. Tox. Chem.*

Westendorf, R.G. (1986) Performance aspects of volatile organics analysis by purge and trap capillary column gas chromatography with flame ionization detectors. Tekmar Technical Papers, Tekmar Company, Cincinnati, Ohio.

Presentations June 1, 1992 - May 31, 1993

Matthews, G.B., R.A. Matthews and W.G. Landis. Multivariate Analyses of Data from Aquatic Toxicity Microcosm Studies: A Comparison of Three Statistical Tests Pacific Northwest SETAC Annual Meeting, Bellingham, WA, June 26-27, 1992.

Landis, W.G., G.B. Matthews, R.A. Matthews and N.J. Shough. Evaluation of the Aquatic Toxicity of the Turbine Fuel Jet-A using Single Species and Microcosm Toxicity Tests. Pacific Northwest SETAC Annual Meeting, Bellingham, WA, June 26-27, 1992.

Landis, W.G., R.V. Sahakian, F.B. Taub and J.H. Taub. Population Effects and Community Dynamics in the Standardized Aquatic Microcosm-modeling Compared to Historical Data Sets. Annual Meeting Society of Environmental Toxicology and Chemistry, November 1992, Cincinnati, OH.

Landis, W. G. , R. A. Matthews, A. J. Markiewicz, N. J. Shough and G. B. Matthews. Evaluation of the Aquatic Toxicity of Two Turbine Fuels Using Microcosm Toxicity Tests. Annual Meeting Society of Environmental Toxicology and Chemistry, November 1992, Cincinnati, OH.

Matthews, G.B., W.G. Landis and R.A. Matthews. Replicability of the Control Group Response in the SAM. Annual Meeting Society of Environmental Toxicology and Chemistry, November 1992, Cincinnati, OH.

Landis, W.G. and A. Sergeant. Incorporation of Multivariate Techniques to the Performance of Ecological Risk Assessments. Annual Meeting Society of Environmental Toxicology and Chemistry, November 1992, Cincinnati, OH.

Landis, W.G., R.A. Matthews and G.B. Matthews. Endpoints: Art or Artifact? Annual Meeting Society of Environmental Toxicology and Chemistry, November 1992, Cincinnati, OH.

Matthews, G.B., W.G. Landis and R.A. Matthews. Nonmetric Clustering and Association Analysis: An Artificial Intelligence Approach to Multispecies Data. Annual Meeting Society of Environmental Toxicology and Chemistry, November 1992, Cincinnati, OH.

Hearne, J. W. and G.B. Matthews. Uncertainty Propagation in Risk Assessment. Annual Meeting Society of Environmental Toxicology and Chemistry, November 1992, Cincinnati, OH.

Landis, W. G., R. A. Matthews and G. B. Matthews. Community level responses to toxicant stress as evaluated by nonmetric clustering and multivariate projections, implications for resource damage and risk assessment. U. S. EPA ERL Corvallis, January 1993, Corvallis, OR.

Landis, W. G., Matthews, R. A. and Matthews, G. B. Ecosystem Dynamics: Wormspace, Chaos and the Implications for Ecological Risk Assessment. USEPA Region 10 Risk Assessment Seminar Series. March 24, 1993, Seattle, WA.

Landis, W. G., Matthews, R. A. and Matthews, G. B. Oscillations detected by multivariate analysis in microcosm toxicity tests with complex toxicants: implications for risk assessment. Society of Environmental Toxicology and Chemistry World Congress, March 30, 1993, Lisbon, Portugal.

Matthews, G. B., W. G. Landis, and R. A. Matthews. Nonmetric clustering and association analysis: risk assessment implications for the interpretation of multispecies toxicity tests and field monitoring. Society of Environmental Toxicology and Chemistry World Congress, March 30, 1993, Lisbon, Portugal.

Grue, C. and W. G. Landis. Indirect effects of contaminants on aquatic wildlife. Society of Environmental Toxicology and Chemistry World Congress, March 30, 1993, Lisbon, Portugal.

Landis, W. G. , R. A. Matthews, G. B. Matthews. Oscillations detected by multivariate analysis in microcosm toxicity tests with complex toxicants: implications for biomonitoring and risk assessment. ASTM Symposium Environmental Toxicology and Risk Assessment , April 27, 1993, Atlanta, GA.

Matthews, G. B., W.G. Landis and R.A. Matthews. Nonmetric clustering and association analysis: implications for the evaluation of multispecies toxicity tests and field monitoring. ASTM Symposium Environmental Toxicology and Risk Assessment , April 27,1993, Atlanta, GA.

Landis, W. G., Matthews, R. A. and Matthews, G. B. Ecosystem Dynamics: Wormspace, Chaos and the Implications for Ecological Risk Assessment. Keynote Address, Ecological Risk Assessment. USEPA Regional Risk Assessment Annual Meeting, May 4, 1993, Atlanta, GA.

Landis, W. G., Matthews, R. A. and Matthews, G. B. Ecosystem Dynamics: Wormspace, Chaos and the Implications for Ecological Risk Assessment. Lecture, Classic Papers in Aquatic Science Course, University of Washington, March 6, 1993, Seattle, WA.

Sandberg, R. S. and W. G. Landis. Use of the Mixed Flask Culture (MFC) Microcosm Protocol to Investigate the Effects of a Pulsed Release of Jet-A Turbine Fuel from Sediments. 1993 Annual Meeting Pacific Northwest Chapter of the Society of Environmental Toxicology and Chemistry, May 20-22, 1993, Newport OR.

Rodgers, S. C. and W. G. Landis. Evaluation of Community Structures Vs Community Function after Exposure to the Turbine Fuel Jet-A. 1993 Annual Meeting Pacific Northwest Chapter of the Society of Environmental Toxicology and Chemistry, May 20-22, 1993, Newport OR.

Markiewicz, A. J., R. A. Matthews, and W. G. Landis. Comparison of the Degradation of Water Soluble Components in Jet Fuel using the Standardized Aquatic Microcosm (SAM) and the Mixed Flask Microcosm (MFC). 1993 Annual Meeting Pacific Northwest Chapter of the Society of Environmental Toxicology and Chemistry, May 20-22, 1993, Newport OR.

Landis, W. G., A. J. Markiewicz, R. A. Matthews and G. B. Matthews. Non-linear Dynamics of Microcosm Experiments after Toxicant Stress Evaluated by Response Volume Projections (Space-Time Worms). 1993 Annual Meeting Pacific Northwest Chapter of the Society of Environmental Toxicology and Chemistry, May 22, 1993, Newport OR.

Scheduled for 1993

Landis, W. G., R. A. Matthews and G. B. Matthews. Use of Novel Methods and the Application of Non-linear Dynamics to the Evaluation of Ecosystem Impacts. Seminar National Oceanic and Atmospheric Administration Hazardous Materials Team, June 21, 1993, Seattle, WA .

Submitted for the 1993 Annual Meeting of the Society for Environmental Toxicology and Chemistry, November 1993, Houston, TX

Landis, W. G. , Matthews, R.A., and Markiewicz, A.J., Matthews, G.B. Comparison of Test Results in the Evaluation of the WSF of Several Jet Fuels Using the Standardized Aquatic Microcosm and the Mixed Flask Culture Protocols. Annual Meeting Society of Environmental Toxicology and Chemistry, November 1993, Houston, TX

Matthews, G. B. , Landis, W.G., and Matthews, R.A.. Response Volumes (Space-time Worms) as a Method for the Visualization of Ecosystem Dynamics and Indirect Effects. Annual Meeting Society of Environmental Toxicology and Chemistry, November 1993, Houston, TX

Landis, W. G., Matthews, R.A., and Matthews, G.B. Non-linear Dynamics of Microcosm Ecosystems and the Inherent Limitations of Risk Assessment. Annual Meeting Society of Environmental Toxicology and Chemistry, November 1993, Houston, TX

Landis, W. G. Matthews, R.A., and Matthews, G.B. Characterization and Classification of Direct and Indirect Effects at the Community and Ecosystem Levels. Annual Meeting Society of Environmental Toxicology and Chemistry, November 1993, Houston, TX

Markiewicz, A. J., Matthews, R.A. and Landis, W.G. Comparison of the Degradation of Water Soluble Components in Jet Fuel Using the Standard Aquatic Microcosm (SAM) and the Mixed Flask Microcosm (MFC). Annual Meeting Society of Environmental Toxicology and Chemistry, November 1993, Houston, TX

Rodgers, S. C. and Landis, W.G. Evaluation of Community Structure and Community Function After Exposure to the Turbine Fuel Jet-A. Annual Meeting Society of Environmental Toxicology and Chemistry, November 1993, Houston, TX

Sandberg, R. S., Roze, M.J., and Landis, W.G. Use of the Mixed Flask Culture (MFC) Microcosm Protocol to Investigate the Effects of a Pulsed Release of Jet-A.. Annual Meeting Society of Environmental Toxicology and Chemistry, November 1993, Houston, TX

Publications in press

Landis, W. G., R. A. Matthews, A.J. Markiewicz, N. A. Shough and G. B. Matthews. *In press.* Multivariate Analyses of the Impacts of the Turbine Fuel Jet-A Using a Microcosm Toxicity Test. *J. Environ. Sci.* Vol 2:2.

Matthews, G. B., R. A. Matthews and W. G. Landis. *In press.* Nonmetric Conceptual Clustering in Ecology and Ecotoxicology. *AI Applications.*

Landis, W. G., R. A. Matthews, A. J. Markiewicz and G. B. Matthews. *In press.* Multivariate Analysis of the Impacts of the Turbine Fuel JP-4 in a Microcosm Toxicity Test with Implications for the Evaluation of Ecosystem Dynamics and Risk Assessment. *Ecotoxicology.*

Submitted Manuscripts

Landis, W. G., G. B. Matthews, R. A. Matthews and A. Sergeant. Application of multivariate techniques to endpoint determination, selection and evaluation in ecological risk assessment. *Env. Tox. Chem.*

Matthews, G. B., R. A. Matthews, W. G. Landis, and J. W. Hearne. Uncertainty propagation in risk assessment. *Env. Tox. Chem.*

Landis, W. G., Matthews, R. A., Markiewicz, A. J. and Matthews, G. B. Non-linear oscillations detected by multivariate analysis in microcosm toxicity tests with complex toxicants: Implications for biomonitoring and risk assessment. *Environmental Toxicology and Risk Assessment-Third Volume, ASTM 1218,* Jane S. Hughes, Gregory R. Biddinger, and Eugene Mones, Eds., American Society for Testing and Materials, Philadelphia, 1994.

Matthews, G. B., R. A. Matthews and W. G. Landis. Nonmetric clustering and association analysis: Implications for the evaluation of multispecies toxicity tests and field monitoring. *Environmental Toxicology and Risk Assessment-Third Volume, ASTM 1218,* Jane S. Hughes, Gregory R. Biddinger, and Eugene Mones, Eds., American Society for Testing and Materials, Philadelphia, 1994.

Students Supported by the Grant and Student Research Projects

Keel, Lester - Anthopluera as a Monitor for the Environmental Impacts of Toxicants (Dr. Landis-Huxley College).

Marklewicz, April J. Fate of Jet Fuel Water Soluble Fraction in the Standardized Aquatic Microcosm (Dr. R. Matthews-Huxley College)

Rodgers, Sara-Comparison of MFC toxicity tests with and without adapted communities (Dr. Landis-Huxley College).

Sahaklan, Robert - Population Dynamics and the Effects of Toxicants on Community Structure (Dr. Landis-Huxley College).

Sandberg, Randall - Modification of the MFC for use in sediment testing (Dr. Landis-Huxley College).

Roze, Michael. Application of RIFFLE program for data evaluation (Dr. G. Matthews-Computer Science).

Professional Collaborators in the Research Program

James W. Hearne, Computer Science Department, Western Washington University

Anne Fairbrother, USEPA Environmental Research Laboratory, Corvallis Oregon

Richard Bennett, USEPA Environmental Research Laboratory, Corvallis Oregon

Anne Sergeant, Office of Research and Development, U.S. EPA, Washington, DC

Frieda B. Taub, School of Fisheries, University of Washington, Seattle

John H. Taub, Seattle, WA.

Interactions and Consultations

Over the last year this research has been translated into technology transfers to DOD and EPA laboratories and the private sector. Apart from presenting the research at national and international meetings, we have been successful in transferring this data and technology during informal meetings or presentations on-site. Below is a list of several of the groups with which we met of transferred information over the last 12 months.

Joseph Dulka, Agricultural Product Department, DuPont Experimental Station, Wilmington, DE. Microcosm use and data analysis.

Lidia Watrud, Team Leader, and Ray Siedler Biotechnology Team, U.S. EPA-Corvallis, OR. Data analysis from terrestrial microcosms.

Nigel Blakley, Department of Ecology, Olympia, WA. Toxicity evaluation of petroleum mixtures.

SETAC Microcosm Workshop. Design and data analysis of microcosms for pesticide evaluations.

ICI Americas. Data analysis of aquatic microcosm studies.

Anne Sergeant, ORD, U.S. EPA., Washington, D.C. Application of multivariate methods to ecological risk assessments.

Heather Gordon, National Research Council, Canada. Riffle program.

Joni A. Torsella, U.S. EPA, Cincinnati, OH. Riffle program.

Patrick A. Thorpe, Grand Valley State University, Allendale, MI. Permtest program.

Charles Hadden, Science Applications International Corp., Oak Ridge, TN. Clustering analysis.

Byron Bodo, Byron A. Bodo & Associates, Canada. nonmetric clustering techniques and Riffle program.

Prof. Hein H. Du Preez, Rand Afrikaans University, South Africa. AI techniques for multispecies toxicity tests.

Scott Ferson, Applied Biomathematics, Setauket, NY. Nonmetric clustering techniques.

APPENDIX A

Publications in Press and Submitted Manuscripts

Multivariate Analyses of the Impacts of the Turbine Fuel Jet-A Using a Standard Aquatic Microcosm Toxicity Test

W. G. Landis*, R. A. Matthews, A. J. Markiewicz, N. J. Shough

Institute of Environmental Toxicology and Chemistry

Huxley College of Environmental Studies

Western Washington University 98225

G. B. Matthews

Computer Science Department

Western Washington University 98225

*To whom correspondence should be addressed

RUNNING HEAD: Multivariate Analyses of Jet-A Toxicity

Key Words: multispecies toxicity testing, Jet-A, microcosm, nonmetric clustering, multivariate analysis

Abstract

Turbine fuels are often the only aviation fuel available in most of the world. Turbine fuels consist of numerous constituents with varying water solubilities, volatilities and toxicities. This study investigates the toxicity of the water soluble fraction (WSF) of Jet-A using the Standard Aquatic Microcosm (SAM). Multivariate analysis of the complex data, including the relatively new method of non-metric clustering, was used and compared to more traditional analyses. The SAM experiment was conducted using concentrations of 0, 1, 5 and 15 percent WSF. The WSF is added on day 7 of the experiments by removing 450 ml from each microcosm including the controls, then adding the appropriate amount of toxicant solution and finally bringing the final volume to 3L with microcosm media.

Analysis of the WSF using purge and trap gas chromatography revealed 55 organic peaks. In the highest WSF concentration treatment group an algal bloom ensued, generated by the apparent toxicity of the WSF of Jet-A to the daphnids. As the test proceeded, the algal populations decreased and were similar to the control values. At the end of the SAM, ostracods exhibited a bloom, with the population density following treatment group in a dose/response manner. Univariate statistics suggested that recovery had taken place by the end of the SAM. Multivariate analysis, however, demonstrated oscillating separations between the 4 treatment groups for the Jet-A experiment. The variables that were most important in distinguishing the four groups

shifted during the course of the 63 day experiment, demonstrating the fallacy of using only one index or only a few measured endpoints in the evaluation of community level interactions.

Introduction

Over the last 15 years a variety of multispecies toxicity tests have been developed with the hope that in doing so, the increased complexity of the test would result in more realistic, community-level responses to the toxicant. However, the addition of more than one species, and the generally longer time periods associated with these multispecies tests, also result in much more complex data sets. Distinguishing toxicant effects from other community-level changes has become one of the most critical obstacles to the interpretation of multispecies data sets.

Multispecies toxicity tests are usually referred to as microcosms or mesocosms, although a clear definition of the size or complexity to distinguish these terms has not been put forth. Multispecies toxicity tests range from approximately 1 L (e. g. , mixed flask cultures) to thousands of liters, as in the case of the pond mesocosms used in pesticide registration testing. The number of species and origin of those taxa can vary widely. In the Standardized Aquatic Microcosm (SAM) (1) developed by Taub and colleagues (2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12) the physical, chemical, and biological components are defined as to species, media and substrate (see Table 1 and Figure 1). In other systems colonization by the importation of sediment or by repeated inoculation forms a natural source is used to establish the model system. Larger systems often use a combination of means to start

Table 1 near here

Figure 1 near here

and maintain a multispecies, interactive community.

One of the major difficulties in the evaluation of multispecies toxicity tests has been the difficulty in the analysis of the large data set on a level consistent with the goals of the toxicity test. Typically, the goals of the toxicity test are:

- to detect changes in the population dynamics of the individual taxa that would not be apparent in single species tests; and,
- to detect community-level differences that are correlated with treatment groups thereby representing a deviation from the control group.

A number of methods have been developed to attempt to satisfy the goals of multispecies toxicity testing. Analysis of variance (ANOVA) is the classical method to examine single variable differences from the control group. However, because multispecies toxicity tests generally run for weeks or even months, there are problems with using conventional ANOVA. These include the increasing likelihood of introducing a Type II error (accepting a false null-hypothesis), temporal dependence of the variables, and the difficulty of graphically representing the data set. Conquest and Taub (13) developed a method to overcome some of the problems by using intervals of non-significant difference (IND). This method corrects for the likelihood of Type II errors and produces intervals that are easily graphed to ease examination. The method is routinely used to examine data from SAM toxicity

tests, and it is applicable to other multivariate toxicity tests. The major drawback is the examination of a single variable at a time over the course of the experiment. While this addresses the first goal in multispecies toxicity testing, listed above, it ignores the second. In many instances, community-level responses are not as straightforward as the classical predator/prey or nutrient limitation dynamics usually picked as examples of single-species responses that represent complex interactions.

Multivariate methods have proved promising as a method of incorporating all of the dimensions of an ecosystem. One of the first methods used in toxicity testing was the calculation of ecosystem strain developed by Kersting (14, 15, 16) for a relatively simple (three species) microcosm. This method has the advantage of using all of the measured parameters of an ecosystem to look for treatment-related differences. At about the same time, Johnson (17, 18) developed a multivariate algorithm using the n-dimensional coordinates of a multivariate data set and the distances between these coordinates as a measure of divergence between treatment groups. Both of these methods have the advantage of examining the ecosystem as a whole rather than by single variables, and can track such proceses as succession, recovery and the deviation of a system due to an anthropogenic input.

However, a major disadvantage of both these methods, and of many conventional multivariate methods, is that all of the data are often incorporated without regard to the units of measurement or the appropriateness of includ-

ing all variables in the analysis. It can be difficult to combine variables such as pH, with units ranging from 0-14, with the numbers of bacterial cells per ml, where low numbers are in the 10^6 range, to say nothing of the conceptual difficulties of adding pH units to counts. Similarly, random variables (i. e., variables with no treatment-related response) indiscriminately incorporated into the analysis may contribute so much noise that they overshadow variables that do show treatment-related effects.

Ideally, a multivariate statistical test used for evaluating complex data sets will have the following characteristics:

- It will not combine counts from dissimilar taxa by means of sums of squares, or other *ad hoc* mathematical techniques, as in the Euclidean and cosine distance measures.
- It will not require transformations of the data, such as normalizing the variance.
- It will works without modification on incomplete data sets.
- It will work without further assumptions on different data types (*e.g.*, species counts or presence/absence data).
- Significance of a taxon to the analysis will not be dependent on the absolute size of its count, so that taxa having a small total variance, such as rare taxa, can compete in importance with common taxa, and

taxa with a large, random variance will not automatically be selected, to the exclusion of others.

- It will provide an integral measure of "how good" the analysis is, *i.e.* whether the data set differs from a random collection of points.

- It will, in some cases, identify a subset of the taxa that serve as reliable indicators of the physical environment.

Recently developed for the analysis of ecological data is a multivariate derivative of artificial intelligence research, nonmetric clustering, that satisfies all these criteria, and has the potential of circumventing many of the problems of conventional multivariate analysis.

In this paper, we use ANOVA and intervals of non-significant difference, and three multivariate techniques to search for meaningful patterns in the data set from a SAM toxicity test using Jet-A turbine fuel. The multivariate techniques include two conventional tests based on the ratio of multivariate metric distances (Euclidean distance and cosine of the vector distance), and one relatively new program, RIFFLE, which employs nonmetric clustering and association analysis (19). All three of the multivariate techniques have proven useful in analyzing complex ecological data sets (20, 21). Of the three, only nonmetric clustering meets all of the criteria listed above (22). The major disadvantage of the RIFFLE program is that, in order to find a clustering of the data points with the desirable qualities listed above,

a massive search through thousands of potential clustering candidates is made before settling on the "right" one. Even after this search, there is no guarantee that RIFFLE finds an optimal clustering. However, in our experience, RIFFLE does find an excellent clustering in reasonable time.

Jet fuels or perhaps more accurately, turbine fuels, are one of the primary fuels for internal combustion engines worldwide and certainly are the most widely available aviation fuel. Over the last 15 years virtually all of the commercial airline operations and charter operations have converted to a turbine engine because of the inherent low operating cost of the power plant, its reliability, and in part to the availability of fuel even in undeveloped areas. In the U. S. military there has been a progressive replacement of conventional piston engine vehicles with turbine equivalents. Standardization on a single type of turbine fuel to relieve logistical demands is also underway. Given the overwhelming predominance of turbine fuel, a fuel spill or accidental release of aviation fuel will likely be one of the prevalent turbine fuels: Jet-A ,used for commercial and general aviation; JP-4, the standard fuel of the U. S. Air Force and Army Aviation; and JP-5, the naval equivalent of JP-4. JP-8 is a new fuel proposed as the standard for all military vehicles using turbine engines.

Along with the environmental considerations, turbine fuels also offer advantages as model complex toxicants for toxicological research. Because of their use as aviation fuel, turbine fuels are produced to stringent specifica-

tions designed to ensure the safety of flight. Therefore, the overall general properties of these materials are tightly controlled. In addition, standard archived samples of the military fuels are maintained for toxicological studies at Wright Patterson, AFB. Jet fuels also tend to be less explosive and also less volatile than gasoline, making the materials easier and safer to use. Like all petroleum products, however, the exact identity of the constituents varies according to the original crude and the refining process.

This paper reports the effects of low concentrations of the water soluble fraction (WSF) of Jet-A on the community incorporated in the SAM. The effects of the WSF on the microcosm communities were subtle. An early increase in algal density was apparent in the treatment groups containing the highest concentrations of the WSF and was matched by a decrease in daphnid populations. Multivariate analysis proved to be more powerful and efficient in highlighting important variables and processes than ANOVA. The variables that were most important is distinguishing treatment-related effects shifted during the course of the experiment. The multivariate analysis also detected oscillations in the similarity of the control and dosed groups that were not apparent using conventional univariate tests. The oscillations may be due to the inherent perturbations in community dynamics, or the effects upon the segments of the community not directly measured, the bacterial detritivores. We discuss the danger of using only one index, or only a few measured endpoints, in the evaluation of community level interactions

in hazard determination and monitoring for risk assessment.

Materials and Methods

Reagents

All chemicals used in the culture of the organisms and in the formulation of the microcosm media were reagent grade or as specified in the protocol (1). Jet-A was provided by Fliteline Services of Bellingham, Washington, U.S.A., and was refined by Chevron. The sample was obtained from the sample valve used for quality control and water sampling to prevent contamination by the refueling apparatus. The shipment lot was recorded and is on file.

Glassware for the preparation of the WSF of Jet-A was washed in non-phosphate soap, rinsed, soaked in 2M HCl for at least 1 h, rinsed ten times with distilled water, dried, and finally autoclaved for 30 min. Microcosm medium, T82MV, acted as the dilutent for the water fraction of the WSF. Twenty-five ml of Jet-A was added to a 2-L separatory funnel, and agitated as follows:

1. Shake separatory funnel for 5 min, releasing built up pressure as necessary.

2. Allow funnel contents to remain undisturbed for 15 min.

3. Shake contents for 5 min, allow to stand 15 min.

4. Continue same pattern for a total time of 2 h.

5. Allow separatory funnel contents to remain undisturbed for 8 h.

At the end of this procedure the mixture was allowed to stand overnight. The next day all but 100 ml of T82MV/WSF mixture from the separatory funnel was drained into a cleaned, sterile 1 L amber glass bottle and capped with a Teflon-lined screw cap. This leaves the lighter, insoluble fuel mixture in the flask. The WSF was used within 24 h or stored at 4° C for no longer than 48 h before use as toxicant mixture.

Gas Chromatography of WSF

The gas chromatography analysis of the WSF used a Tekmar LSC 2000 Purge and Trap (P&T) concentrator system in tandem with a Hewlett Packard 5890A Gas Chromatograph and a Flame Ionization Detector (FID) (23, 24, 25). Instrument blanks and deionized distilled water blanks were used to verify the P&T and GC columns cleanliness prior to analysis of the WSF samples. A five ml sample was injected into a 5 ml sparger, purged with pre-purified nitrogen gas for 11 min and dry purged for 4 min. Volatile hydrocarbons, purged from the sample and collected on the Tenax/Silica Gel column, were desorbed at 180° C directly onto the gas chromatograph SPB-5, 30m x 0.53 mm ID 1.5 μm film, fused silica capillary column. The column, at 35° C, was held at that temperature for 2 min, increased to 225°

C at 12 °C/min and held at that temperature for 5 min. A Spectra-Physics 4290 Integrator was used to record the FID signal output of the volatile hydrocarbons that were separated and eluted from the column by molecular weight. A comparison was then made of the sample chromatograph to n-paraffin and n-naphtha chromatograph standards, prepared and analyzed under the same conditions. A summary of the specification for the P&T gas chromatography used for this experiment is listed in Table 2. Table 2 near here

Algal and Daphnic Toxicity Tests

In order to determine the appropriate WSF concentrations to use for for the SAM microcosm, a series of short-term toxicity tests were performed. These included 96 h algal growth inhibition tests using three species of algae and a 48 h *Daphnia magna* toxicity test.

Algal growth inhibition

Algal growth inhibition tests were performed to determine the toxicity of the WSF of the various fuels using *Chlamydamonas reinhardii*, *Ankistrodesmus falcatus* and *Selenastrum capricornutum*.

The test algae were grown in a semi-flow through culture apparatus on the microcosm media T82MV and taken during log phase growth for inoculation into the test flasks. Five hundred ml Erlenmeyer flasks with ground glass stoppers were used as test chambers. Each test chamber contained a

total of 100 ml of the control or treatment solution. Two replicates of of the following dilutions were used: 0.0, 6.25, 12.5, 25, 50 and 100 percent WSF. All dilutions of the WSF were made using T82MV. The test organisms were added at a concentration of approximately 3.0×10^4 cells/ml. Test mixtures were incubated at 20.0° C ± 1.0° C with a 12:12 h light/dark cycle. Cell densities were determined every 24 h during the 96 h test using a Newbauer Counting Chamber.

The cell numbers were then plotted against the WSF concentrations. If possible, a least-squares regression line was drawn and the IC_{50} (the concentration at which algal growth is reduced to 50% of the control) was determined. An ANOVA was used to determine if any of the groups were significantly different.

D. magna toxicity test

Daphnia magna acute toxicity tests (26) were conducted using T82MV medium at concentrations of 0, 6.25, 12.5, 25, 50 and 100 percent WSF. Ten neonates were placed in 250 ml beakers containing 200 ml of test solution, with two replicates at each concentration. After 24 and 48 h, the number of dead were recorded. Data were analyzed graphically and statistically to obtain an estimate of the EC_{50}.

SAM Protocol

The 64-day SAM protocol follows most of the procedures described in (1). Table 1 describes the organisms, conditions and modifications used for the Jet-A experiment. The microcosms consist of 4 L glass jars containing 3 L of sterile T82MV microcosm medium and autoclaved sediment consisting of 200 g silica sand and 0.5 g of ground chitin. The sediment is autoclaved in the experimental jar immersed in a water bath to a point above the sand and chiten level during sterilization. This procedure helps prevent breakage of the jars and subsequent loss of replication. The microcosms were inoculated with ten algal, four invertebrate, and one bacterial species. The microcosms were incubated at 20.0° C ± 1.0° C, with illumination set at 79.2 μEm^{-2} sec^{-1}, PhAR ranging from 78.6-80.4, and a 16/8 day/night cycle. The numbers of organisms, dissolved oxygen (DO) and pH were determined twice weekly.

The major modification on the SAM protocol was the means of toxicant delivery. The test material was added on day 7 by stirring each microcosm, removing 450 ml from each container, and then adding appropriate amounts of the WSF to produce concentrations of 0, 1.5, and 15 percent WSF. After toxicant addition the final volume was adjusted to 3L. No attempt was made to filter and retain the organisms withdrawn during the removal of the 450 ml prior to toxicant. All graphs and statistical analysis start with the next

mpling day, day 11.

Data Analysis

All data were recorded onto standard computer entry forms and checked for accuracy. The parameters that were calculated included the numerical densities for each of the species, DO, DO gain and loss, net photosynthesis/respiration ratio (P/R), pH, algal species diversity, algal biovolume, and biovolume of available algae (1). For each of the parameters, the IND was determined (13). The INDs and the average values for each treatment group were plotted against time to identify significant differences between the treatments and control. Note that algal biovolume, algal species diversity, and available algae are all derived variables based on the algal counts. The P/R ratio was derived using daytime and nighttime oxygen concentrations.

Three multivariate significance tests were used. Two of them were based on the ratio of multivariate metric distances within treatment groups *vs* between treatment groups. One of these was calculated using Euclidean distance and the other with cosine of vectors distance (27, 28). The third test used nonmetric clustering and association analysis (19).

The biotic parameters used for our multivariate analysis of the SAM data are listed in Table 3. Treating a sample on a given day as a vector of values, $\vec{x} = \langle x_1 \ldots x_{17} \rangle$, with one value for each of the measured biotic parameters,

Table 3 near here.

allows multivariate distance functions to be computed. Euclidean distance between two sample points $\vec{x}$ and $\vec{y}$ was computed as:

$$\sqrt{\sum_i (x_i - y_i)^2}$$

The cosine of the vector distance between the points $\vec{x}$ and $\vec{y}$ was computed as:

$$1 - \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum y_i^2}}$$

Subtracting the cosine from one yields a distance measure, rather than a similarity measure, with the measure increasing as the points get farther from each other.

The within-between ratio test used a complete matrix of point-to-point distance (either Euclidean or cosine) values. For each sampling date, one sample point $\vec{x}$ was obtained from each of six replicates in the four treatment groups, giving a 24×24 matrix of distances. After the distances were computed, the ratio of the average within group distance (W) to the average between group distance (B) was computed (W/B). If the points in a given treatment group are closer to each other, on average, than they are to points in a different treatment group, then this ratio will be small. The significance of the ratio is estimated with an approximate randomization test (29). This test is based on the fact that, under the null hypothesis, assignment of points to treatment groups is equivalent to a random assignment, the treatment having no effect. The test, accordingly, randomly assigns the 24 points to

(pseudo) groups, and recomputes the W/B ratio, a large number of times (500 in our tests). If the null hypothesis is false, this randomly derived ratio will be larger, on average, than the the W/B ratio obtained from the actual treatment groups. By taking a large number of random reassignments, a valid estimate of the probability under the null hypothesis is obtained as $(n+1)/(500+1)$, where n is the number of times a ratio less than or equal to the actual ratio was obtained (29).

In the clustering association test, the data were first clustered independently of treatment group, using nonmetric clustering and the computer program RIFFLE (22). Because the clustering analysis is naive to treatment group, the clusters may, or may not correspond to treatment effects. Under the null hypothesis, there should be no correspondence between the clustering and the treatment groups. To evaluate whether the clusters were related to the treatment groups, the association between clusters and treatment groups was measured in a 4×4 contingency table, each point in treatment group i and cluster j being counted as a point in frequency cell ij. Significance of the association in the table was then measured with Pearson's χ^2 test (30), defined as

$$\chi^2 = \sum_{i,j} \frac{(N_{ij} - n_{ij})^2}{n_{ij}}$$

where N_{ij} is the actual cell count and n_{ij} is the expected cell frequency,

obtained from the row and column marginal totals N_{+j} and N_{i+} as

$$n_{ij} = \frac{N_{+j}N_{i+}}{N}$$

where $N = 24$ is the total cell count. The significance (probability) for this value of χ^2 was computed using a standard procedure (31).

Results

GC Analysis

Originally, 55 peaks were distinguishable as constituents of the WSF derived from Jet-A (Figure 2). At the end of the 63 day course of the experiment, and using the same method, virtually all of the peaks had disappeared from the water column.

Figure 2 near here

Short Term Toxicity Tests

Three sets of 96 h algal toxicity tests were performed (using *A. falcatus*, *S. capricornutum*, and *C. reinhardii*). None of the tests demonstrated dramatic toxicity or enhancement under the test conditions. *Selenastrum* demonstrated a trend towards a slight enhancement of growth, but not in any dose response manner (Figure 3a). *Ankistrodesmus* seems to indicate a slight inhibition, but not in a traditional dose/response manner (Figure 3b). No difference was observed in the *Chlamydamonas* toxicity tests, likely due to the slow growth of this strain under these test conditions.

Figure 3 near here

The 48 h *D. magna* toxicity tests did demonstrate an acute toxicity resulting in a graphically derived EC_{50} of approximately 10 percent WSF. Therefore, we expected that the highest concentration in the SAM experiments would adversely impact the daphnid populations shortly after the toxicant addition.

Univariate ANOVA and IND results

Algae The largest increase in algal population density occurred in Treatment 4 (see Figure 4). The peak density is approximately four times that of the control replicates at day 21. Treatment 3 also exhibited an early increase in algal density during the first fourteen days after the introduction of the toxicant. The algal densities in the control and lowest treatment group both exhibited decreases in algal densities during the same period. At the end of the experiment the total algal numbers are not significantly different although Treatments 3 and 4 are consistently lower. Algal species diversity also generally declined in each of the treatment groups but not in relationship to dose.

Figure 4 near here

Daphnia The control and lowest treatment group demonstrated similar patterns of daphnid population dynamics (Figures 5a and 5b). The early increases in the algal densities in the two highest treatment groups are likely due to the inhibition of reproduction and the survival of the neonates in the

period after dosing. In Treatment 3 we saw an increase in the number of small daphnids and the overall population starting on day 14 (Figure 5c). Treatment 4 did not show a major increase in the daphnid populations until day 17; the peak was not reached until after day 30 (Figure 5d). Figure 5 near here

Ostracods At the end of the experiment the average population density in the control treatments was approximately twice that of Treatment 4 (Figure 6). The population densities in the other treatments were ranked in a dose/response manner. The ranking was consistent from day 49 onward. The IND plot does not pick any of these results as being significantly different from the control. Figure 6 near here

Philodina and Protozoans The hypotrichous protozoa were present only in low densities throughout the experiment. *Philodina* did not appear in appreciable numbers until after day 35 in any of the treatments. Although the control harbored the lowest density at the end of experiment, compared to Treatments 3 and 4, the IND plots did not show any significant differences (Figure 7). The difficulty in sampling rapidly growing and declining populations with asynchronous growth is apparent. Although trends may be suggested, conventional analysis did not see a significant effect. Figure 7 near here

pH and P/R ratio The P/R ratio, measured by changes in oxygen concentration, exhibited a dose response relationship early in the experiment

with Treatments 3 and 4 being significantly different from the controls according to the IND plots (Figure 8a). Excursions from the control appear to occur on day 53 but again, this may be a chance event.

pH responded in a dose response manner to the addition of Jet-A. During the period of the algal blooms pH was significantly higher than in the two highest treatment groups than in the control, as determined by the IND plots (Figure 8b). On day 49 a deviation from the control in a dose/response manner was detected. However, with the multiple comparisons being made it is difficult to attribute such an event to the treatment. At the end of the experiment all of the groups resembled the control.

Figure 8 near here

Multivariate results

The significance levels for the three multivariate tests performed for each sampling day are graphed in Figure 9. All tests agree that a significant difference between treatment groups was observed through day 25. From day 28 to day 39, the effect diminished until there were no significant effects observable. However, significant effects were again observable from day 46 through day 56, after which they again disappeared for days 60 and 63.

Figure 9 near here

In Figure 10, the average cosine distances within the control group and between the control group and each of the three treatment groups are plotted on a log scale. The initial, strong effect, from day 11 to day 25, is easily seen as a large distance from Treatments 1 (control) and 2, together, to

Figure 10 near here

both Treatment groups 3 and 4. Group 3 subsequently moves closer to the control. The period of no significant difference, from day 35 to day 46, is also clear. During the second period of significant difference, from day 49 to 59, a perfect dose-response relationship for all three treatments is seen, with higher doses becoming more distant from the control. This dose-response relationship is consistently maintained over a period of eleven days, for four sampling dates, days 49, 53, 56, and 59. In general, a dose-response relationship like this was not observed earlier, although the magnitude of the distances was considerably greater.

Also of interest are the variables that best described the clusters and the stability of the importance of the variables during the course of the experiment. Table 4 lists the variables determined to be important in determining the clusters, ranked by importance, for each sampling day as determined by nonmetric clustering. In general, the number of variables that were important was larger during the start of the test, and lower at the end. In addition, a great deal of variabillity in rankings is apparent during the course of the SAM. The number of sampling dates when a variable was deemed important in cluster formation is listed in Table 5. *Ankistrodesmus* was the most consistent of the variables, being ranked in 12 out of the 16 sampling dates. Medium *Daphnia* was also ranked often. However, variables like Ostracod and *Philodina* did not become important until later in the experiment.

Table 4 near here

Table 5 near here

Discussion

Our examination of individual parameters provided only a limited, and somewhat distorted view of the SAM microcosm response to Jet-A. The univariate data analysis did indeed show that there were some significant responses to the toxicant by individual taxa and chemistry; however, the responses were scattered over time, and did not present a logical, coherent pattern. Futhermore, the individual responses we could detect were typically gross aberrations of the microcosm, signifying wild swings in a taxon's population density over time. If you kill or restrict the production of most of the *Daphnia*, the next microcosm response is likely to be an algal bloom. Measuring these types of gross responses to the toxicant do not provide much more insight into the fate of the toxicant in the ecosystem than do the short-term single-species tests.

However, the multivariate analysis reveals a much more interesting dynamic. Although not particularly toxic in the short term toxicity testing, Jet-A had detectable effects upon the dynamics of the multispecies test system, effects which persisted until the end of the experiment. It is important to note that the original WSF mixture was no longer present at the end of the SAM experiment, no doubt lost to volatilization or biotransformation and biodegradation by the biota.

Extrapolation from a simple system to precise estimates of risk to aquatic

systems is a process filled with scientific uncertainty. However, the initial imbalance in predator/prey dynamics and the apparent oscillation of even a simple system, point to effects not observable using single species toxicity testing. The repeated divergence of the dosed replicates from the controls can be accounted for in two basic ways:

- It might reflect the functioning of the community in terms of parameters not directly sampled by the SAM protocol.
- It might be a persistent fluctuation in community structure initiated by the initial stress, but is only periodically visible, as if it were an incompletely dampened oscillation in the systems.

We will now briefly consider each of these.

The multivariate statistics suggest a complex pattern of multiple divergences and convergences in the similarities between treatment groups. Much as an ecosystem could be expected to display the rise and fall of species assemblages, the SAM microcosms appear to indicate that the first divergence is only the beginning of a series of responses. Using nonmetric clustering, we were able to list the variables that were the most important for separating the treatment group clusters for each day that measurements were collected (see Tables 4 and 5). The list of variables suggests that the first divergence, which occurred from about day 11 through day 32, results from predator/prey interactions between primary producers (algae) and first

order consumers (*Daphnia*). Theoretically, this divergence should be characterized by the following properties:

- The divergence will be fast, because the algae and *Daphnia* populations are introduced into the microcosm after being cultured in optimal laboratory conditions, in artificially high densities, and therefore are unstable. Predation, or the lack of predation, will cause rapid changes in the algal densities of prey species.

- The divergence will be short-lived, because the populations are unstable in the nutrient-rich early successional microcosm. There will be a tendency for the microcosms to drift away from the early "treatment" effect into a more stable community based on both algae and detritus as the food source for the secondary consumers. Initially, this drift may mask treatment effects and be interpreted as recovery of the system.

The first divergence is the only type of response that is normally searched for in microcosm tests using conventional statistics. This response is typical of many reported SAM experiments (9, 10, 32, 33).

The second divergence occurred from about day 46 through day 60. During this time, *Daphnia* and some of the algal taxa were often still important in the cluster development; however, other secondary consumers (Ostracods and *Philodina*) entered the list. The second divergence therefore may rep-

resent the long-term effects of the initial toxicant on a more successionally mature community that is fueled by both algae and detritus. If so, the second divergence should have the following characteristics:

- It has been strongly influenced by detritus quality. Detritus is conditioned by bacteria and fungi, which are highly sensitive to toxins but unmeasured in the microcosm. Also, detritus that has passed through the gut of a consumer (e.g., consumed algae) is different from detritus that originates directly from dead algae (unconsumed). Therefore, the quality of the detritus may be highly affected by the treatment, but none of the factors influencing the effects will be measured directly.

- Secondary consumers of detritus and bacteria are no less affected by the quality of their food source than algal consumers, so the treatment-related alterations of the quality of detritus and bacteria will cause differences in the secondary consumer populations.

- Therefore, the second divergence may still represent a direct response to the initial treatment effects, but because it occurs late in the microcosm experiment and is difficult to detect with univariate statistics, it is easily misinterpreted as noise or the effects of a degredation product.

A study of the detritus and bacteria present in late successional microcosms may answer these questions.

However, an alternative explanation may also explain the second divergence, without invoking direct impacts of unseen biotic components of the system. The initial perturbation may be spread through the system, and persist continuously through the experiment, while the convergence seen during the middle of the experiment may be an observational artifact. In effect, the systems may be moving in different directions and simply pass by each other during certain time intervals. As the various groups converge and then reseparate, the second divergence may be seen as a separate event, but in fact this separation is a continuation of the dynamics initiated earlier. The illusion of recovery may simply be a momentary, accidental, confluence. It may well be the case that not every divergence from the control treatment has an observable cause directly related to it in time; differentiating these effects from those due to unobserved consumers, detritus, degradation products or other population and community dynamics is challenging.

Another important characteristic of this experiment is the dynamics of the variables characterized as important by the multivariate analysis. Taken separately, none of the biotic variables used by the multivariate analysis could clearly point to the second departure from the control group response, although hints and suggestions abounded. The sampling variance was simply too high, especially in the protozoa, rotifers and ostracods. However, when correlations were taken on a replicate-by-replicate basis using multivariate analysis, the trends were clear and statistically significant. Even pH,

a variable with a low sampling error, could not clearly distinguish the second divergence, although the IND did show a significant difference late in the expeirment. Without corroboration, the points outside the IND could be considered outliers, improbable events. However, the multivariate analysis demonstrated a clear and significant dose/response relationship. Nonmetric clustering was also able to select the variables that were important in distinguishing the four treatment groups, although the variables contributing to the differentiation changed from sampling day to sampling day.

These data suggest that reliance upon any one variable, or an index of variables, probably would have missed the second oscillation of the treatments. The implications are important. Currently, only small sections of ecosystems are monitored, or a heavy reliance is placed upon so-called indicator species. These data suggest that to do so is dangerous and may produce misleading interpretations resulting in costly errors in management and regulatory judgements.

Several questions raised by this experiment are now the goals of future research. The dynamics of the loss of jet fuels from the SAM systems is currently being investigated in greater depth. The multiphased response seen in this experiment may have been a chance event. Additional testing of related jet fuels is also currently being conducted. The implications for hazard and risk assessment are also significant and we are investigating the incorporation of multivariate analysis into these processes. Finally, the

effects of size and community structure abound. The SAM system is relatively simple. Data sets incorporating more diverse species assemblages and of varying sizes are being investigated for comparison.

In summary, we can make the following observations: The water soluble fraction of Jet-A has a low toxicity to algae but a greater toxicity to the cladoceran *D. magna*. In the microcosm study, only some of the effects of Jet-A can be attributed to differential toxicity. At least two oscillations from control are distinguishable in the treatment group responses. Multivariate analysis is crucial in observing effects with highly dimensioned and typically noisy data sets. Multivariate analysis points to the dynamic nature of variables important in distinguishing treatment groups. Reliance upon indices that condense data, or upon indicator species, may be misleading in determining effects of stressors upon biological communities.

Acknowledgement

This research is supported by USAFOSR Grant No. AFOSR-91-2091 DEF. Robert V. Sahakian and Lester W. Keel participated in the preparation and sampling of the microcosms. Mark V. Haley, U.S. Army Chemical Research Division and Engineering Center kindly supplied the original cultures used in these experiments.

References

1. ASTM E1366-91. (1991): Standard practice for standardized aquatic microcosm: fresh water, volume 11.04. American Society for Testing and Materials, Philadelphia.

2. F. B. Taub. (1969): Gnotobiotic models of freshwater communities. Verh Internat. Verein. Limnol., 17:485-496.

3. F. B. Taub. (1976): Demonstration of pollution effects in aquatic microcosms. Intern J. Environmental Studies., 10:23-33.

4. F. B. Taub and M. E. Crow. (1978): Loss of a critical species in a model (laboratory) ecosystem. Verh. Internat. Verein. Limnol., pp. 1270-1276.

5. M. E. Crow and F. B. Taub. (1979): Designing a microcosm bioassay to detect ecosystem level effects. Intern. J. Environmental Studies, pp. 141-147.

6. F. B. Taub, M. E. Crow, and H. J. Hartmann. (1980): Responses of aquatic microcosms to acute mortality. In Jr. Giesy, J. P., Ed.: Microcosms in Ecological Research, pp. 513-535. Technical Information Center, U. S. Department of Energy, Washington, D. C.

7. A. C. Kindig, L. C. Loveday, and F. B. Taub. (1983): Differential sensitivity of new versus mature synthetic microcosms to streptomycin sulfate treatment. In W. E. Bishop, R. D. Cardwell, and B. B. Heidolph, Eds.: Aquatic Toxicology and Hazard Assessment: Sixth Symposium. ASTM 802, pp. 192-203. American Society for Testing and Materials, Philadelphia.

8. F. B. Taub and P. L. Read. (1983): Standardized Aquatic Microcosm Protocol: Final Report Contract No. 223-80-2352, Vol II. Food and Drug Administration, Washington, D. C..

9. F. B. Taub, A. C. Kindig, and L. L. Conquest. (1987): Interlaboratory testing of a standardized aquatic microcosm. In W. J. Adams, G. A. Chapman, and W. G. Landis, Eds.: Aquatic Toxicology and Hazard Assessment: 10th Volume, ASTM STP 971, pp. 385-405. American Society for Testing and Materials, Philadelphia.

10. F. B. Taub, A. C. Kindig, L. L. Conquest, and J. P. Meador. (1988): Results of the interlaboratory testing of the standardized aquatic microcosm protocol. In G. Suter and M. Lewis, Eds.: Aquatic Toxicology and Hazard Assessment: Eleventh Symposium ASTM. American Society for Testing and Materials, Philadelphia.

11. F. B. Taub. (1988): Standardized aquatic microcosm - development and testing. Aquatic Ecotoxicology, II.

12. F. B. Taub. (1989): Standardized aquatic microcosms. Environm. Sci. Technol. 23(9):1064-1066.

13. L. L. Conquest and F. B. Taub. (1989): Repeatability and reproducibility of the standard aquatic microcosm: Statistical properties. In U. M. Cowgill and L. R. Williams, Eds.: Aquatic Toxicology and Hazard Assessment: 12th Volume, ASTM STP 1027, pp. 159-177. American Society for Testing and Materials, Philadelphia.

14. K. Kersting. (1984): Development and use of an aquatic micro-ecosystem as a test system for toxic substances. Properties of an aquatic micro-ecosystem IV. Int. Revue ges. Hydrobiol., 69(4):567-607.

15. K. Kersting. (1985): Properties of an aquatic micro-ecosystem v. ten years of observations of the prototype. Verh. Internat. Verein. Limnol., 22:3040-3045.

16. K. Kersting. (1988): Normalized ecosystem strain in micro-ecosystems using different sets of state variables. Verh. Internat. Verein. Limnol., 23:1641-1646.

17. A. R. Johnson. (1988): Diagnostic variables as predictors of ecological risk. Environmental Management, 12(4):515-523.

18. A. R. Johnson. (1988): Evaluating ecosystem response to toxicant stress: a state space approach. In W. J. Adams, G.A. Chapman, and W.G. Landis, Eds.: Aquatic Toxicology and Hazard Assessment: 10th Volume, ASTM STP 971, pp. 275-285. American Society for Testing and Materials, Philadelphia.

19. G. B. Matthews and R. A. Matthews. (1991): A model for describing community change. In Pesticides in Natural Systems: How Can Their Effects Be Monitored? Proceedings of the Conference, Environmental Research Laboratory/ORD, Corvallis Oregon. EPA 9109/9-91011.

20. G. B. Matthews, R. A. Matthews, and B. Hachmoller. (1991): Mathematical analysis of temporal and spatial trends in the benthic macroinvertebrate communities of a small stream. Canadian Journal of Fisheries and Aquatic Sciences, 48(11):2184-2190.

21. R. A. Matthews, G. B. Matthews, and William J. Ehinger. (1991): Classification and ordination of limnological data: a comparison of analytical tools. Ecological Modelling, 53:167-187.

22. G. B. Matthews and J. W. Hearne. (1991): Clustering without a metric. IEEE Transactions on Pattern Analysis and Machine Intelligence,

13(2):175-184.

23. ASTM D3710. (1986): Boiling range distribution of gasoline and gasoline fractions by gas chromatography. Annual Book of ASTM Standards, 5.03:99-113.

24. ASTM D2887. (1986): Boiling range distribution of petroleum fractions by gas chromatography. Annual Book of ASTM Standards, 5.02:658-667.

25. ASTM D-2 Proposal (P167). (1989): Manual on hydrocarbon analysis 4th Ed. Series MNL3. American Society for Testing and Materials, Philadelphia.

26. ASTM E729-88a. (1988): Standard guide for conducting acute toxicity tests with fishes, macroinvertebrates, and amphibians, volume 11.04. American Society for Testing and Materials, Philadelphia.

27. I. J. Good. (1982): An index of separateness of clusters and a permutation test for its significance. J. statist. comp. Simul., 15:81-84.

28. E. P. Smith, K. W. Pontasch, and J. Cairns, Jr. (1990): Community similarity and the analysis of multispecies environmental data: a unified statistical approach. Water Res., 24(4):507-514.

29. E. W. Noreen. (1989): Computer Intensive Methods for Testing Hypotheses. Wiley-Interscience, New York, NY.

30. S. E. Fienberg. (1985): The Analysis of Cross-Classified Categorical Data. Mit Press, Cambridge, Ma.

31. W. H. Press, B. P. Flannery, A. A. Teukolsky, and W. T. Vetterline. (1990): Numerical Recipes in C, the Art of Scientific Computing. Cambridge University Press, New York, NY.

32. W. G. Landis, N. A. Chester, M. V. Haley, D. W. Johnson, W. T. Muse, Jr., and R. M. Tauber. (1989): The utility of the standard aquatic microcosm as a standard method for ecotoxicological evaluation. In G. Suter and M. Adams, Eds.: Aquatic Toxicology and Environmental Fate: Eleventh Volume ASTM STP -1007, pp. 353-367. American Society for Testing and Materials, Philadelphia.

33. W. G. Landis, M. V. Haley, and N. A. Chester. (In press): The use of the standardized aquatic microcosm in the evaluation of degradative bacteria in reducing impacts to aquatic ecosystems. In W. G. Landis, J. Hughes, and M. Lewis, Eds.: Environmental Toxicology and Risk Assessment: First Volume ASTM STP-1179. American Society for Testing and Materials, Philadelphia.

Table 1. Summary of Test Conditions for Conducting SAM Jet-A

Organisms	
Organisms per chamber:	Algae (added on Day 0 at initial concentration of 10^3 cells for each algae species): *Anabaena cylindrica*, *Ankistrodesmus* sp., *Chlamydomonas reinhardi* 90, *Chlorella vulgaris*, *Lyngbya* sp., *Scenedesmus obliquus*, *Selenastrum capricornutum*, *Stigeoclonium* sp., and *Ulothrix* sp. Animals (added on Day 4 at the initial numbers indicated in parentheses): *Daphnia magna* (16/microcosm), *Cypridopsis* sp. (ostracod) (6/microcosm), Hypotrichs [protozoa] (0.1/mL), and *Philodina* sp. (rotifer) (0.03/mL)
Experimental design	
Test vessel type and size:	One-gallon (3.8 L) glass jars16.0 cm wide at the shoulder, 25 cm tall with 10.6 cm openings
Medium volume:	3000 mL added to each container
Number of replicates x concentrations:	6x4
Reinoculation:	Once per week add one drop (circa 0.05 mL) to each microcosm from a mix of the ten species = 5 x 10^2 cells of each alga added per microcosm
Addition of test materials:	Test material added day 7 by removing 450 mL from each container and then adding appropriate amounts of the WSF to produce concentrations of 0, 1, 5 and 15 percent WSF. After toxicant addition the final volume was adjusted to 3L.
Sampling frequency:	2 times each week
Test duration:	63 days
Physical and chemical parameters	
Temperature:	20 to 25°C
Light intensity:	80 µE m^{-2} photosynthetically active radiation s^{-1} (850 to 1000 fc)
Photoperiod:	12 h light / 12 h dark
Medium:	Medium T82MV
Sediment:	Composed of silica sand (200 g), ground, crude chitin (0.5), and cellulose powder (0.5 g) added to each container
Measurements:	Algal, invertebrate and protozoa counts, pH, dissolved oxygen, optical density, Parameters calculated included the concentrations of each of the species, DO, DO gain and loss, net photosynthesis/respiration ratio (P/R), pH, algal species diversity, daphnid fecundity, algal biovolume, and biovolume of available algae.

Table 2. Purge and Trap and Gas Chromatograph Specifications for the Analysis of Jet-A Water Soluble Fraction

Tekmar LSC 2000 Purge and Trap column and conditions:

Sample size: 5 mL
Valve, mount and line initial temperature: 30°C
Purge pressure: 140 kPa
Purge: 11 minutes at 42.6 cm/sec. N_2
Dry purge time: 4 minutes
Trap: Tenax/Silica Gel, 1/8" x 12", SS
Desorb preheat temperature: 175°C
Desorb temperature and time: 180°C for 4 minutes
Bake temperature and time: 180°C for 5 minutes

Hewlett Packard 5890A Gas Chromatograph column and conditions:

Column head pressure: 30 kPa
Carrier Gas: Nitrogen, Flow rate: 46.1 cm/sec.
Hydrogen flow rate: 40 cm/sec. Air flow rate: 350 cm/sec.
Column temperature program: 35°C/2 min. // 12°C/min. to 225°C/5 min.
Detector: Flame Ionization Detector
Integrator: Spectra-Physics 4290

Table 3. Biotic parameters used in the multivariate statistical tests. Biotic variables such as diversity, available biovolume, and total algal biovolume are not used since they are derived from and therefore not independent of the variables listed above.

- Anabaena
- Ankistrodesmus
- Chlamydomonas
- Chlorella
- Daphnia
  - Ephipia
  - Small Daphnia
  - Medium Daphnia
  - Large Daphnia
- Hypotricha
- Lyngbya
- Miscellaneous sp.
- Ostracod (Cyprinotus)
- Philodina (Rotifer)
- Scenedesmus
- Selanastrum
- Stigeoclonium
- Ulothrix

Table 4. Important variables as determined by Non-metric clustering ranked according to contribution for each sampling day. Some variables such as *Ankistrodesmus* were consistently important in determining group clusters throughout the experiment. Some of the variables such as **Ostracod** and Philodina were more important in the latter stages of the experiment. Note that the order of importance of even the more common contributors often changed from day to day, with no one variable being consistently ranked, *Ankistrodesmus* being the closest.

Day	Important Variables in Determining Clusters in Rank Order
11	M. Daphnia, Chlorella, Chlamydamonas, Ulothrix, S. Daphnia,Selanastrum,Scenedesmus
14	S. Daphnia, M. Daphnia-Selenastrum[1], Chlamydamonas, Chlorella, L. Daphnia, *Ankistrodesmus*
18	*Ankistrodesmus*, S. Daphnia, Chlorella, Chlamydamonas, Selanstrum, L. Daphnia
21	*Ankistrodesmus*, S. Daphnia, L. Daphnia-M. Daphnia, Scenedesmus
25	Scenedesmus, S. Daphnia, L. Daphnia, Chlorella, Philodina-M. Daphnia
28	*Ankistrodesmus*, L. Daphnia, Scenedesmus
32	S. Daphnia, M. Daphnia, *Ankistrodesmus*, Chlorella
35	*Ankistrodesmus*
39	M. Daphnia-Selenastrum, **Ostracod**-*Ankistrodesmus*
42	M. Daphnia, **Ostracod**, Scenedesmus
46	Scenedesmus, *Ankistrodesmus*, S. Daphnia. M. Daphnia
49	Chlorella, Philodina, *Ankistrodesmus*, Lyngbya
53	*Ankistrodesmus*, **Ostracod**, Chlorella
56	M. Daphnia-Scenedesmus, *Ankistrodesmus*, Lyngbya
60	Lyngbya, M. Daphnia, Philodina, Chlorella
63	Chlorella, *Ankistrodesmus*, Philodina, **Ostracod**

[1] Hyphen between variables denotes equal rank

Table 5. Variable According to Success in Determining Clusters as Defined by Non-metric Clustering. Variables such as *Ankistrodesmus* and the *Daphnia* classes were important in the course of this study. However, reliance on any particular organism or a small combination would have poorly described the dynamics of the system.

Variable	Ranked
Ankistrodesmus	12
M. Daphnia	11
Chlorella	9
Scenedesmus	7
S. Daphnia	6
L. Daphnia	5
Ostracod	4
Philodina	4
Selenastrum	4
Lyngbya	3
Ulothrix	1

Figures

Figure 1. Timeline for the Standardized Aquatic Microcosm Jet-A Experiment. Each step of this 63 day protocol is choreographed according to ASTM E 1366-91. The modifications to the protocol are the elimination of *Nitchia* and *Hyalella azteca* and the modification of the method for toxicant delivery.

Figure 2. Trap and Purge Gas Chromatography Results for the WSF of Jet-A. Originally 55 peaks are distinguishable as constituents of the WSF derived from Jet-A. At the end of the 63 day course of the experiment and using the same method virtually all of the peaks have disappeared from the water column.

Figure 3. 96 h Algal Toxicity Tests. Toxicity tests were performed with *A. falcatus, S. capricornutum* and *C. reinhardi*. None of the tests demonstrated dramatic results. Selenastrum demonstrated a trend towards a slight enhancement of growth, but not in any dose response manner (Figure 3a.). *A. falcatus* seems to indicate a slight inhibition, but not in a traditional dose response manner (Figure 3b.). No difference was observed in *C. reinhardi* toxicity tests, likely due to the slow growth of this strain under these test conditions.

Figure 4. Patterns in Algal Communities. The algal densities in the control and lowest treatment group both exhibited decreases in algal densities until day 21 (Figures 4a and 4b). Treatment 3 (Figure 4c) exhibited an increase in algal density during the first fourteen days after the introduction of the toxicant. The largest increase in algal population density occurred in Treatment 4 (Figure 4d). The peak density is approximately four times that of the control replicates at day 21. At the end of the experiment the total algal numbers are not significantly different although Treatments 3 and 4 are consistently lower.

Figure 5. Daphnid Population Dynamics. The control and lowest treatment group demonstrated similar patterns of daphnid population dynamics (Figures 5a and 5b). The early increases in the algal densities in the two highest treatment groups are likely due to the inhibition of reproduction and the survival of the neonates in the period after dosing. In Treatment 3 day 14 first saw an increase in the number of small daphnids and the overall population (Figure 5c). Treatment 4 did not see a major increase in the daphnid populations until day 17, and the peak, the highest of the treatment groups, was not reached until over midway through the experiment (Figure 5d).

Figure 6. Ostracod Population Dynamics. The average population density in the control treatments is approximatly twice that of Treatment 4, the highest concentration. In between the populations densities

are ranked in a dose response manner. Although suggestive and not readily apparent in the other biological data, the apparent dose response falls within the IND plot surrounding the control.

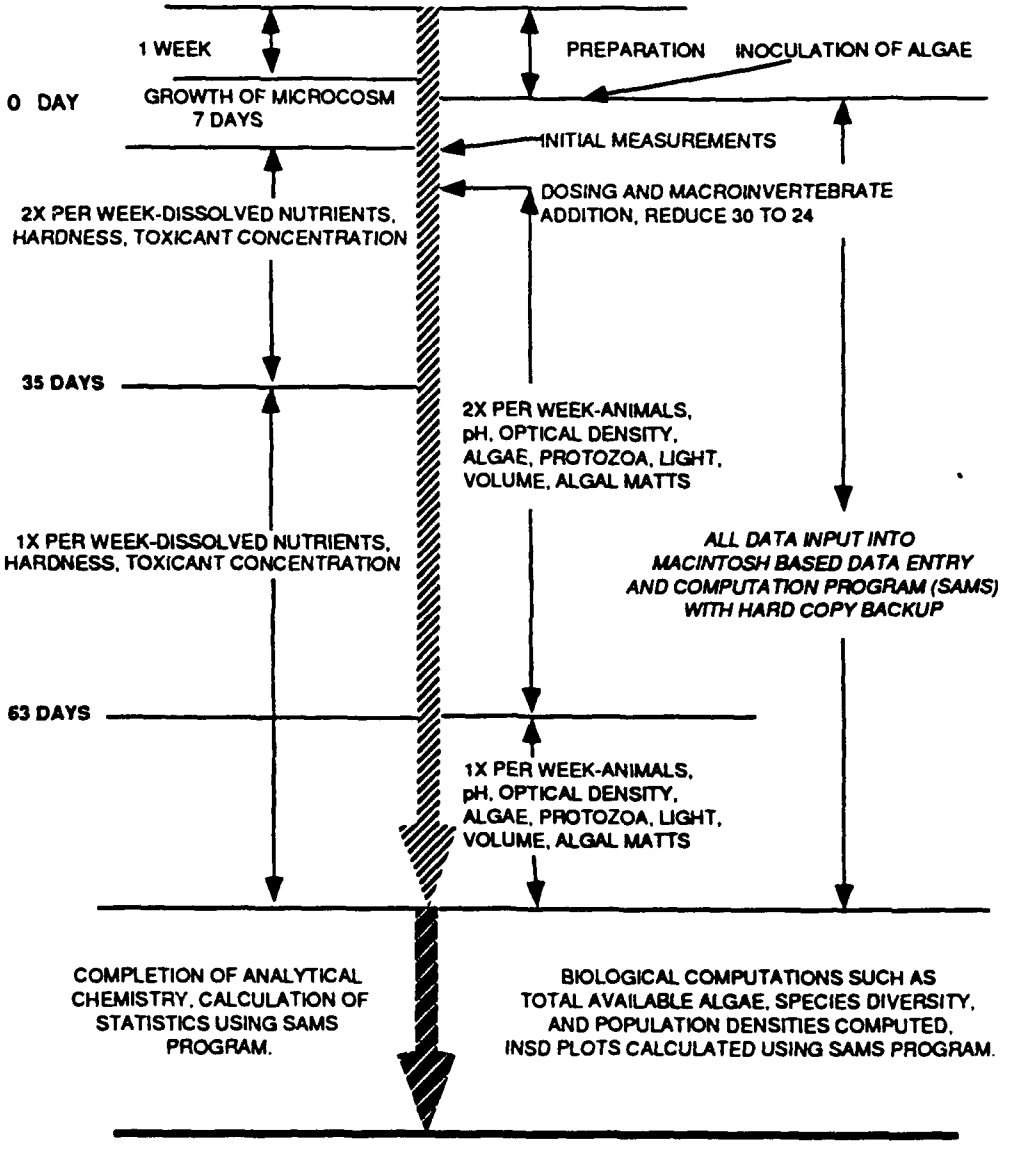
Figure 7. Philodina Population Dynamics. The population dynamics of the Philodina suggest a treatment effect towards the end of the experiment. As with the ostracods the sampling error is too large to distinguish such an effect using conventional univariate techniques. The bars are standard deviations for the means of each sampling day.

Figure 8. Photosynthesis/Respiration ratio and pH. As with the biological data, the chemical data detect a dramatic early effect but do not clearly indicate other deviations from the control occurring later in the test. The photosynthesis/respiration ratio (Figure 8a) clearly illustrates an effect during the early segment of the experiment. On day 53 one of the treatment groups exceeds the IND but by itself this could be classified as a rare event, not truly statistically significant. Again pH (Figure 8b) demonstrates the early deviation and suggests a late effect as the treatment groups exceed the IND.

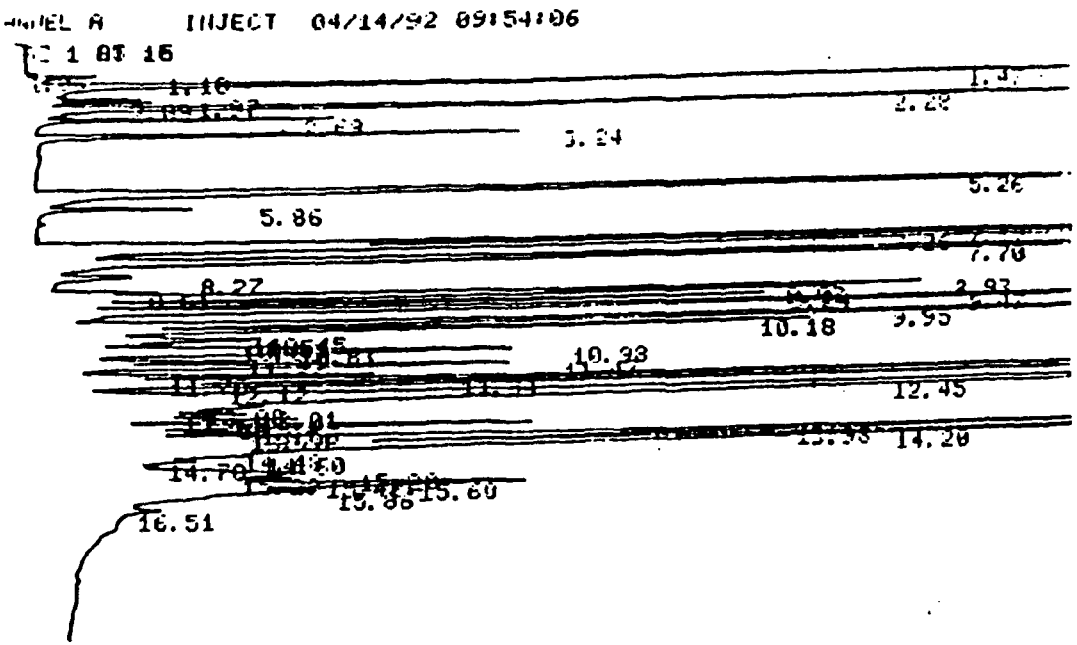
Figure 9. Significance levels of the three multivariate statistical tests for each sampling day. Note that there are two periods, early and late ones, where the clustering into treatment groups is significant at the 95 percent confidence level or above.

Figure 10. Cosine distance from the control group to each of the treatments for each sampling day. Note that large differences are apparent early in the SAM. During the middle part of the 63 day experiment the distances between the replicates of Treatment 1, the control group, is as large as the distances to the treatment groups. However, later in the experiment the distances from the dosed microcosms to the control again increase.

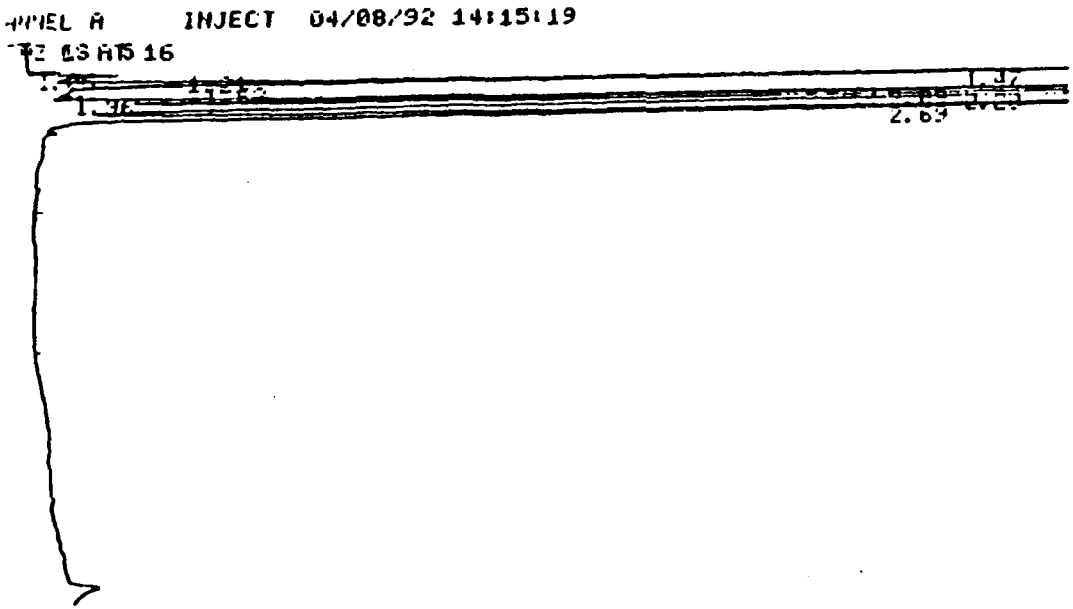
Time Line-Standardized Aquatic Microcosm

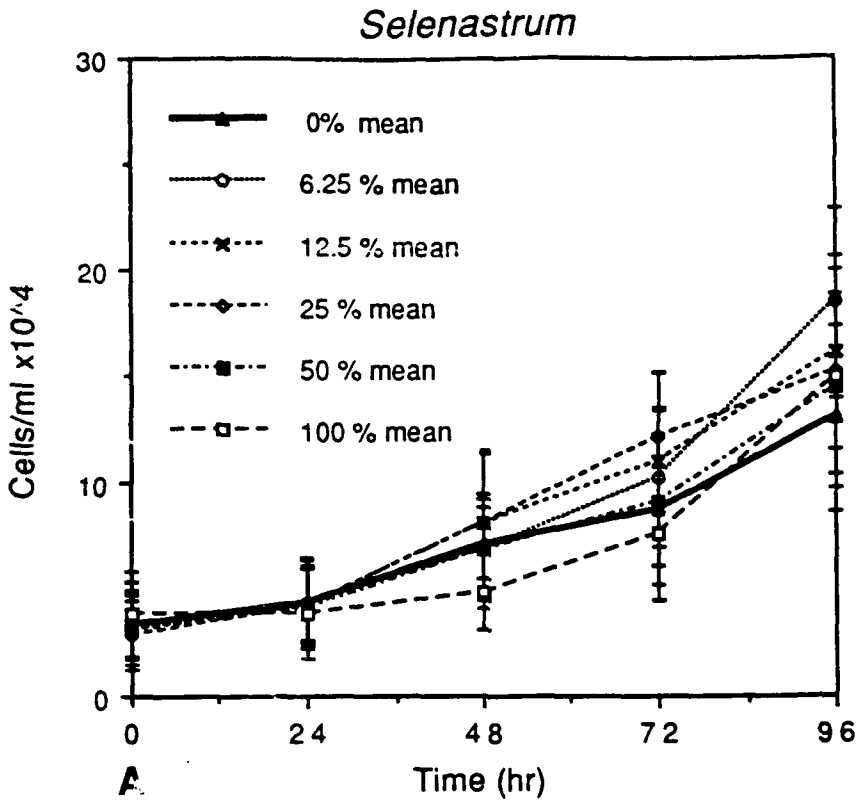


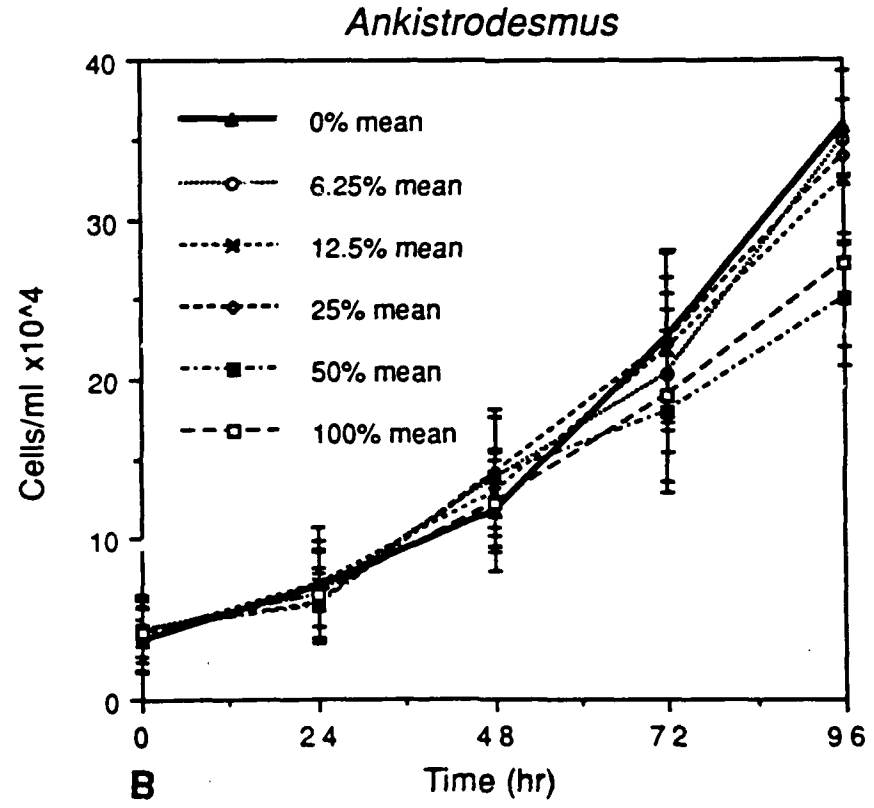
A

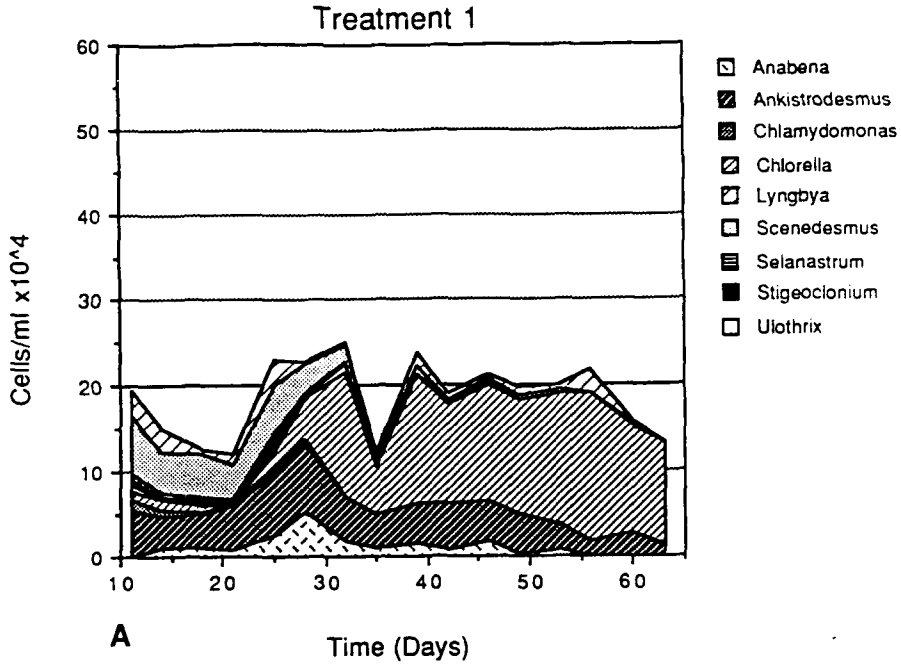

INJECT 04/14/92 09:54:06
5.86
10.18
10.33
12.45
14.20
15.60
16.51

B

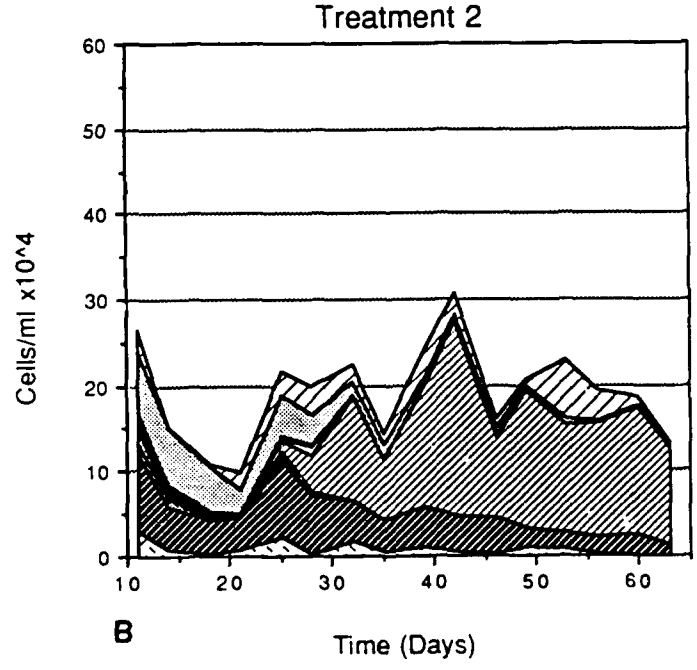

INJECT 04/08/92 14:15:19


Selenastrum
30
20
10
0
Cells/ml x10^4
0% mean
6.25 % mean
12.5 % mean
25 % mean
50 % mean
100 % mean
0
24
48
72
96
Time (hr)
A

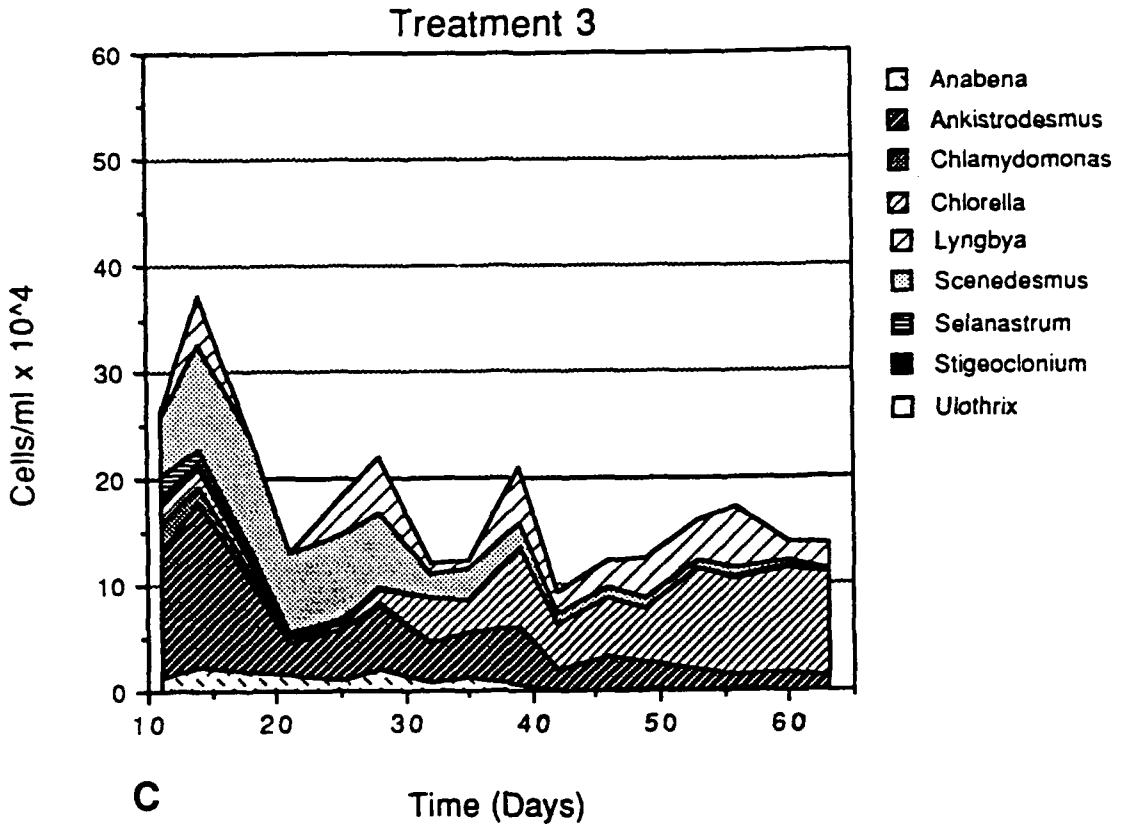

Ankistrodesmus
40
30
20
10
0
Cells/ml x10^4
0% mean
6.25% mean
12.5% mean
25% mean
50% mean
100% mean
0
24
48
72
96
Time (hr)
B


Treatment 1
Cells/ml x10^4
60
50
40
30
20
10
0
10
20
30
40
50
60
Time (Days)
Anabena
Ankistrodesmus
Chlamydomonas
Chlorella
Lyngbya
Scenedesmus
Selanastrum
Stigeoclonium
Ulothrix

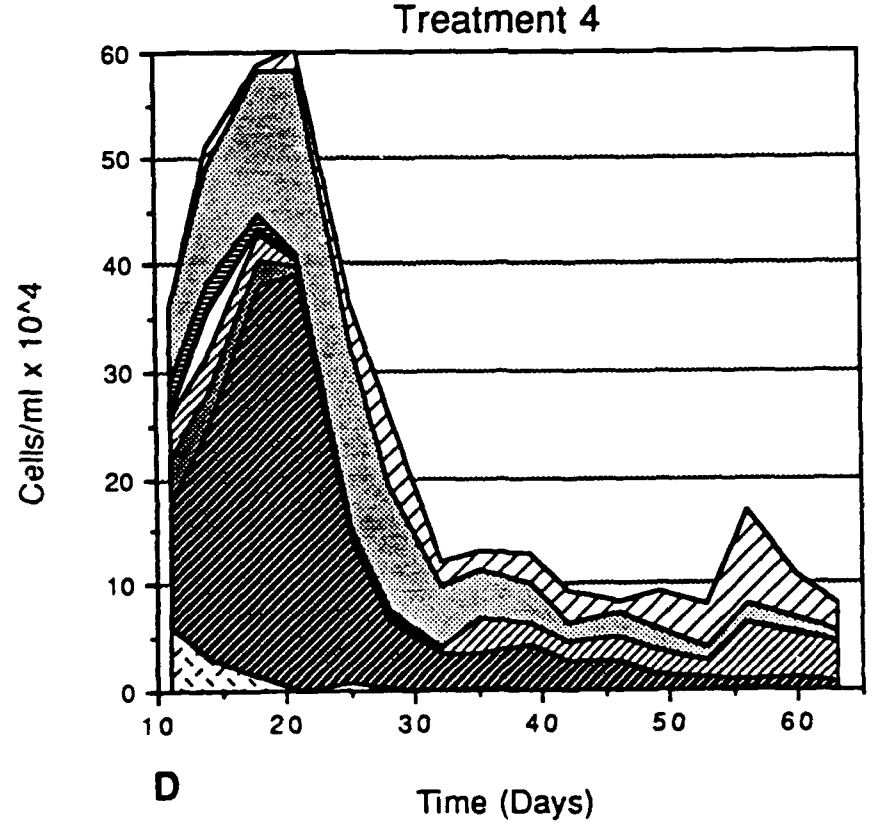
A


Treatment 2
Cells/ml x10^4
60
50
40
30
20
10
0
10
20
30
40
50
60
Time (Days)

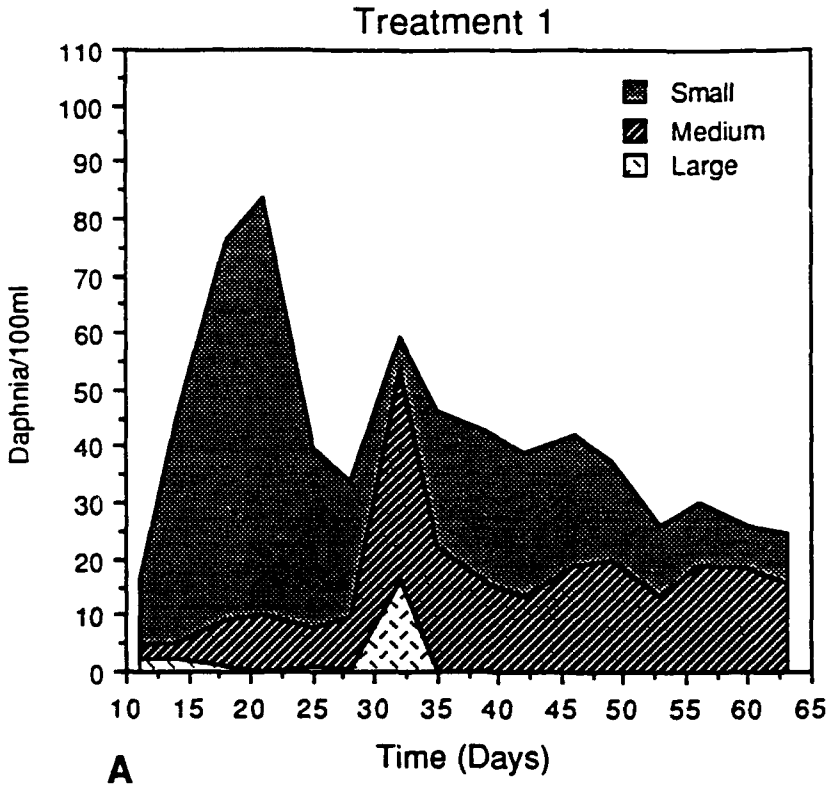
B

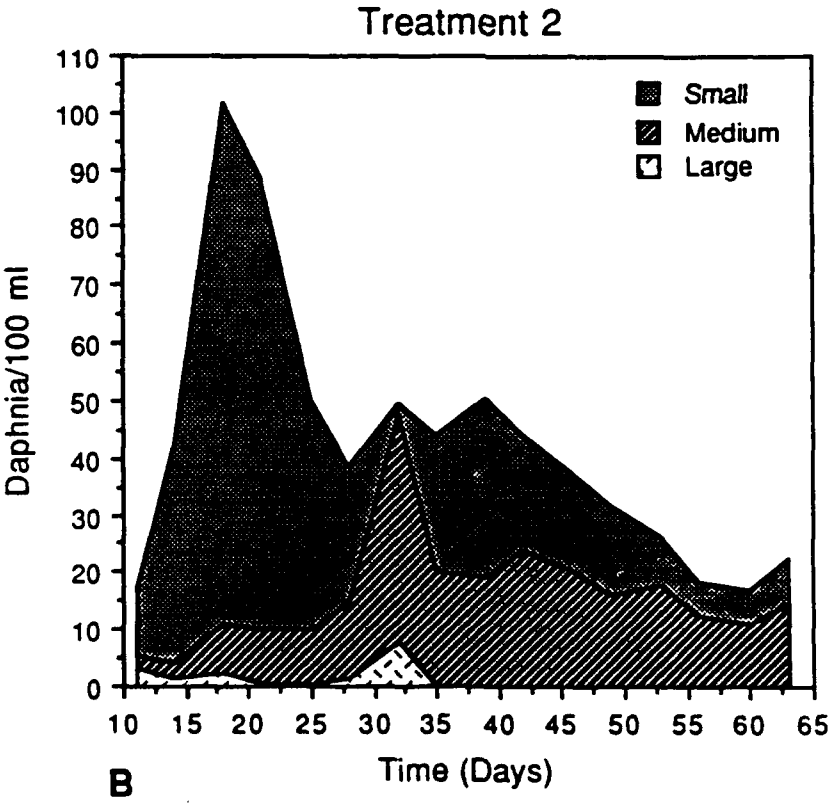

Treatment 3
Cells/ml x 10^4
60
50
40
30
20
10
0
10
20
30
40
50
60
Time (Days)
Anabena
Ankistrodesmus
Chlamydomonas
Chlorella
Lyngbya
Scenedesmus
Selanastrum
Stigeoclonium
Ulothrix

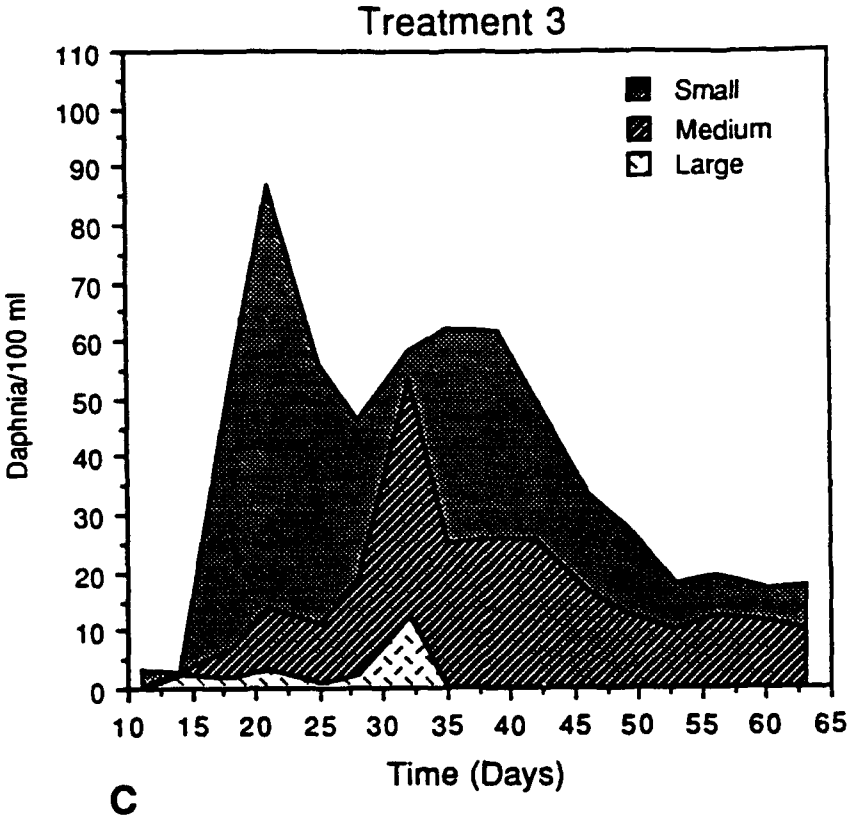
C


Treatment 4
Cells/ml x 10^4
60
50
40
30
20
10
0
10
20
30
40
50
60
Time (Days)

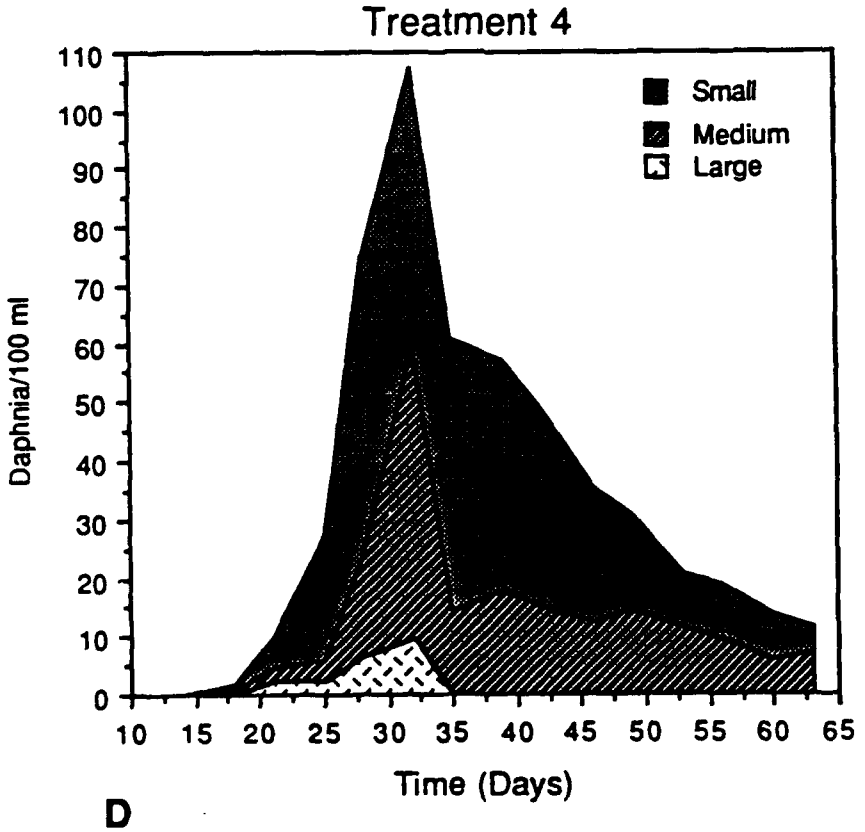
D


Treatment 1
Small
Medium
Large
Daphnia/100ml
110
100
90
80
70
60
50
40
30
20
10
0
10 15 20 25 30 35 40 45 50 55 60 65
Time (Days)
A

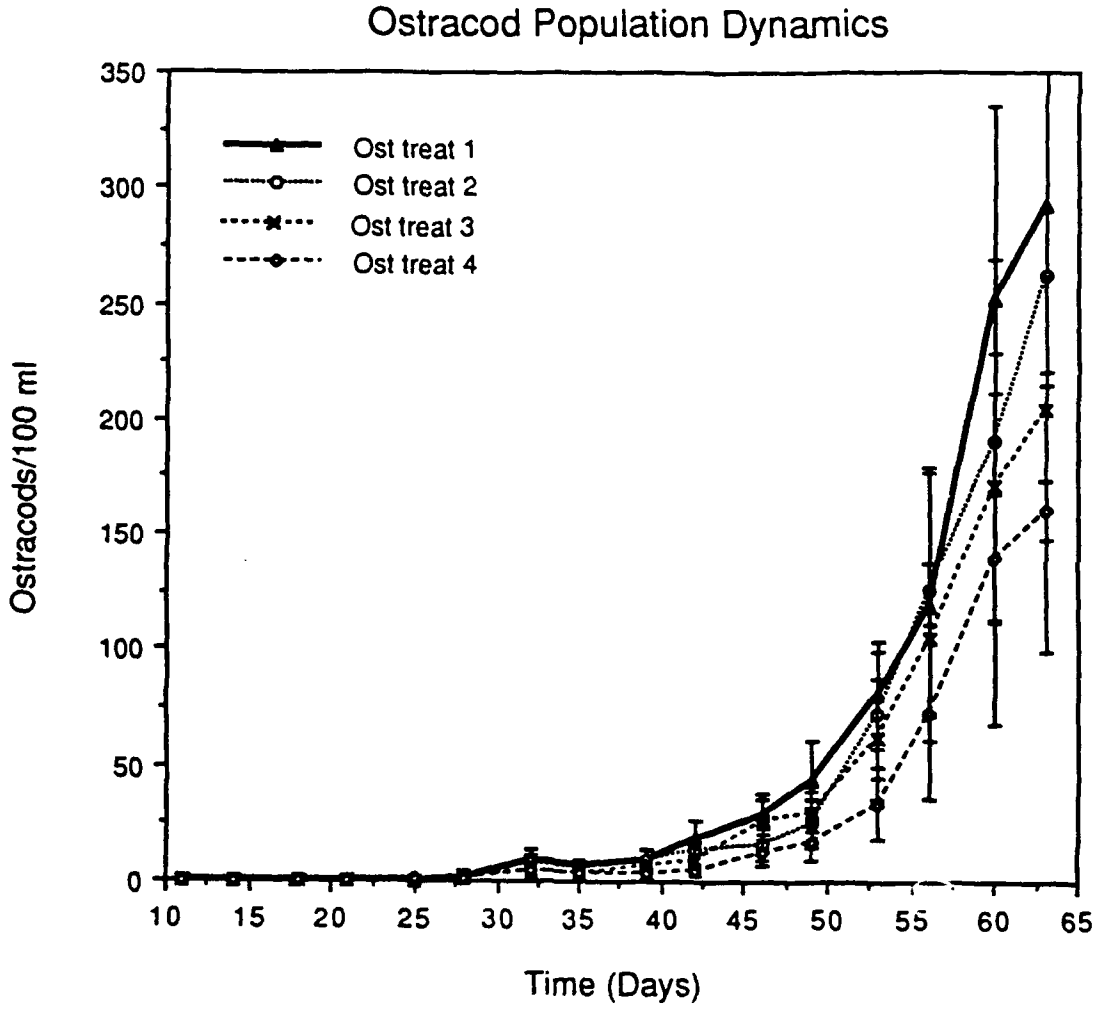

Treatment 2
Small
Medium
Large
Daphnia/100 ml
110
100
90
80
70
60
50
40
30
20
10
0
10 15 20 25 30 35 40 45 50 55 60 65
Time (Days)
B

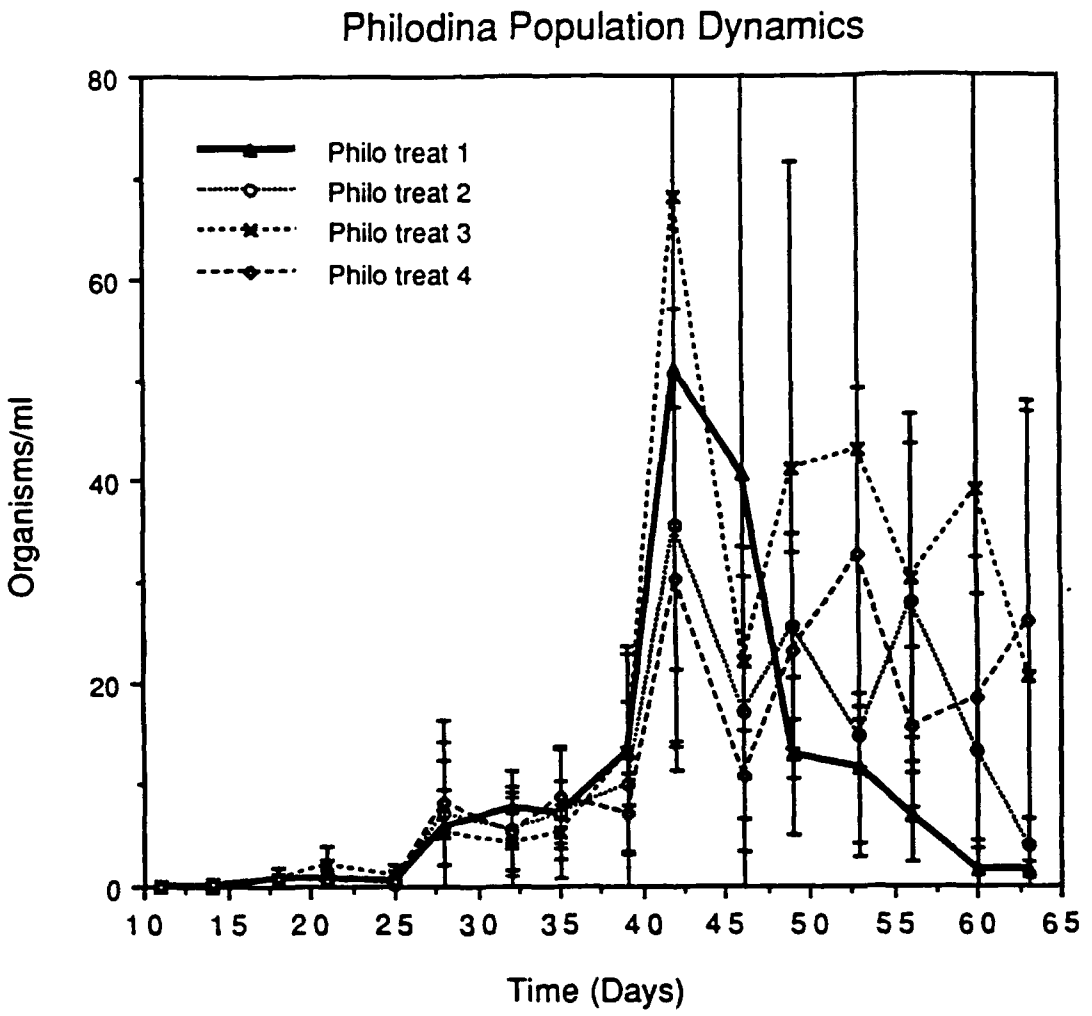

Treatment 3
Small
Medium
Large
Daphnia/100 ml
0
10
20
30
40
50
60
70
80
90
100
110
10 15 20 25 30 35 40 45 50 55 60 65
Time (Days)

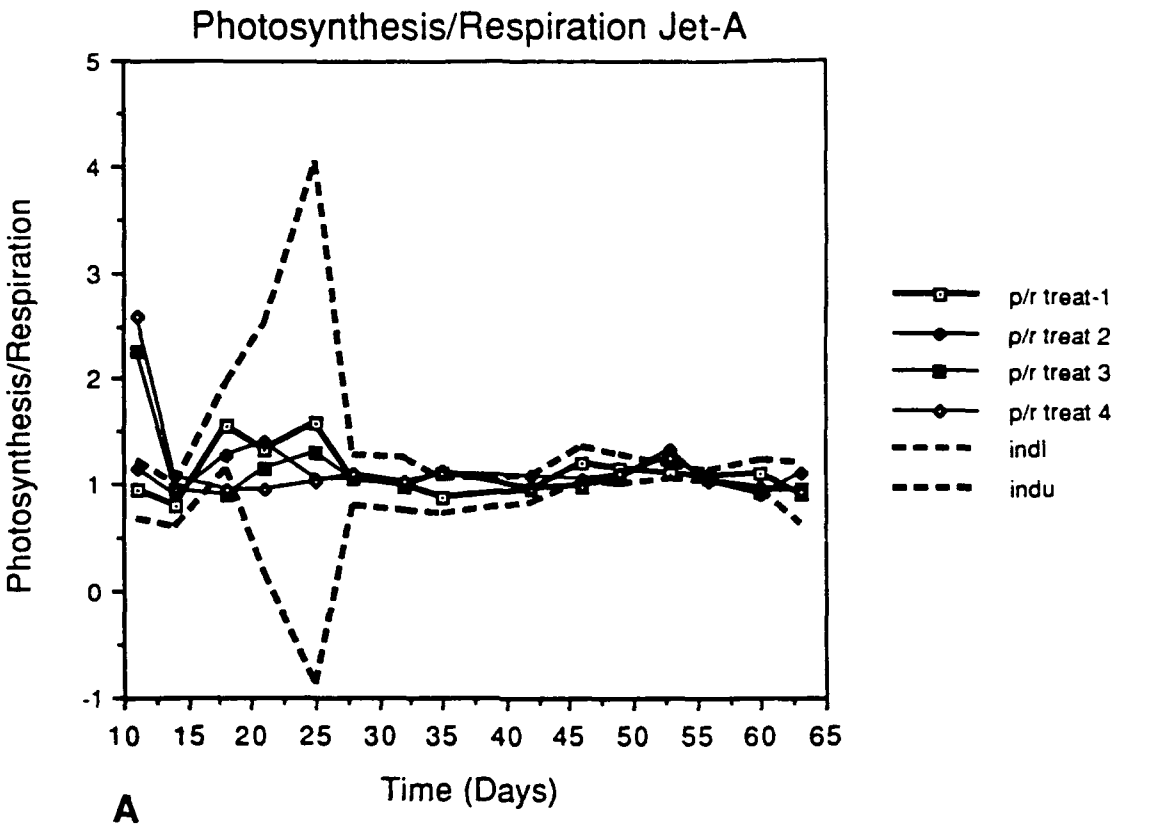
C


Treatment 4
Small
Medium
Large
Daphnia/100 ml
0
10
20
30
40
50
60
70
80
90
100
110
10 15 20 25 30 35 40 45 50 55 60 65
Time (Days)

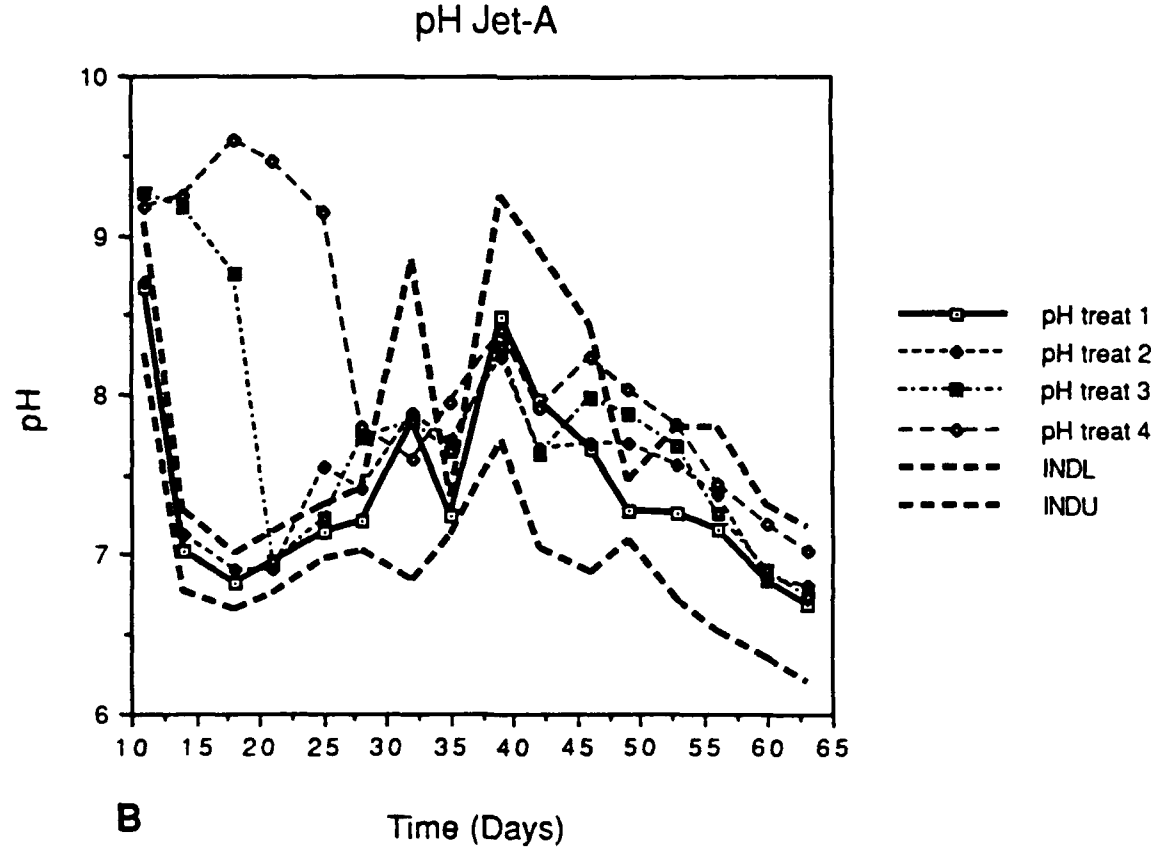
D


Ostracod Population Dynamics
Ost treat 1
Ost treat 2
Ost treat 3
Ost treat 4
Ostracods/100 ml
Time (Days)
0
50
100
150
200
250
300
350
10
15
20
25
30
35
40
45
50
55
60
65

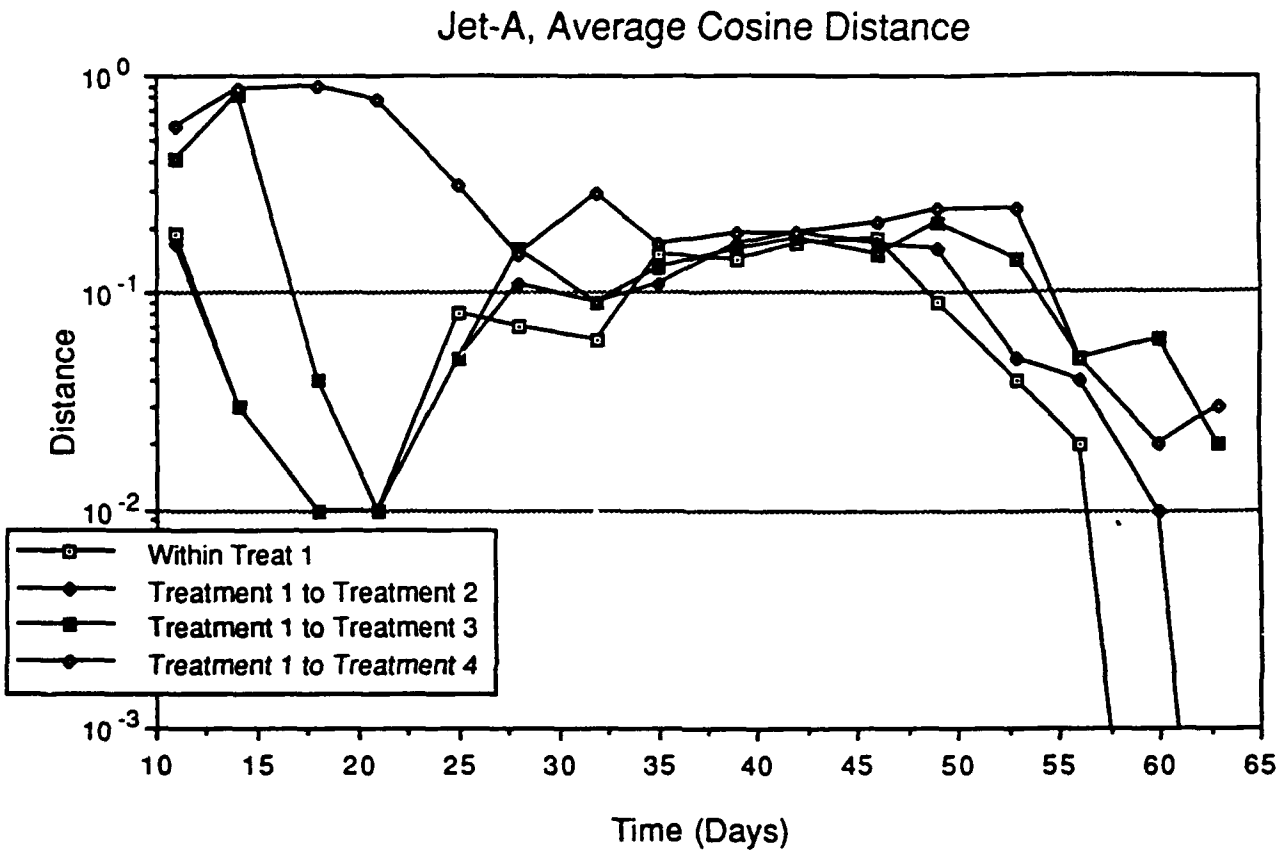

Philodina Population Dynamics
Philo treat 1
Philo treat 2
Philo treat 3
Philo treat 4
Organisms/ml
0
20
40
60
80
10
15
20
25
30
35
40
45
50
55
60
65
Time (Days)

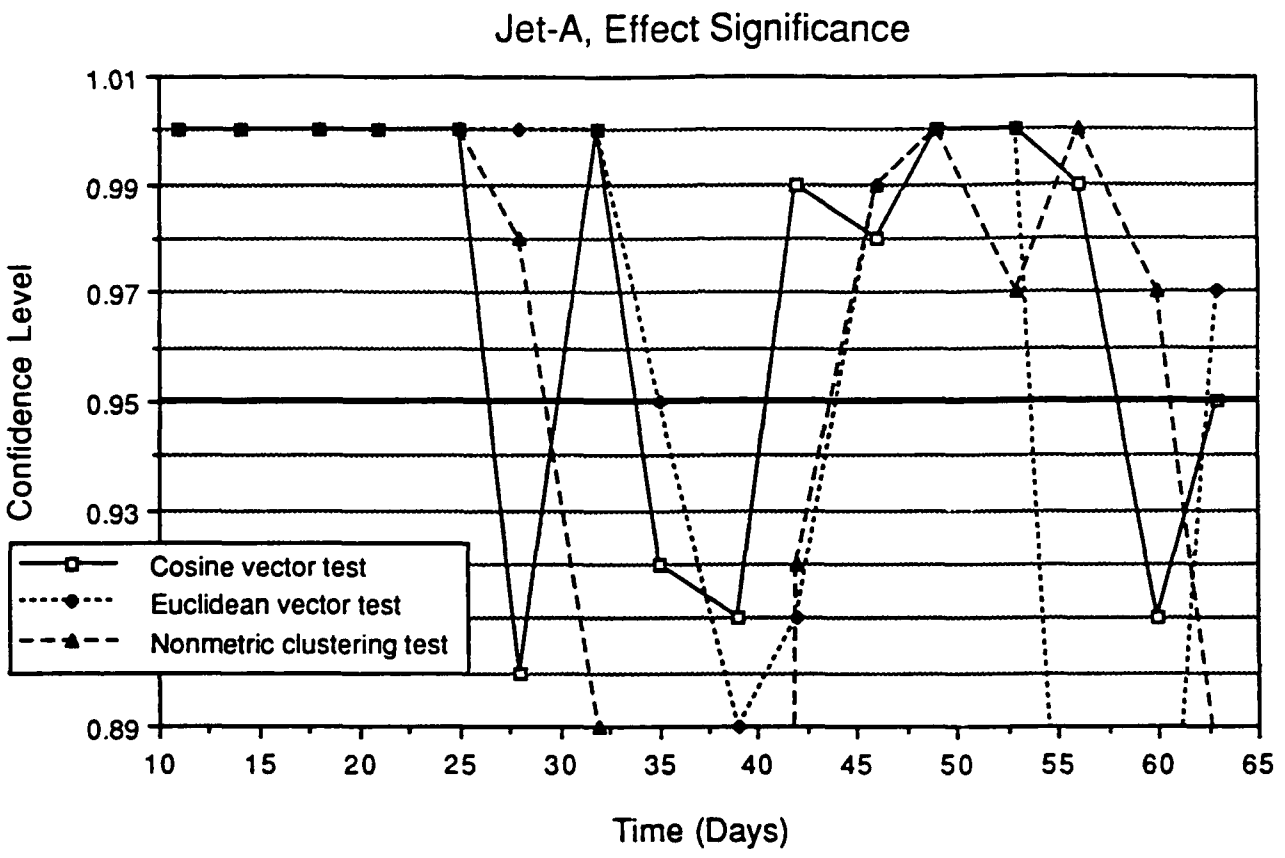

Photosynthesis/Respiration Jet-A
Photosynthesis/Respiration
5
4
3
2
1
0
-1
10 15 20 25 30 35 40 45 50 55 60 65
Time (Days)
p/r treat-1
p/r treat 2
p/r treat 3
p/r treat 4
indl
indu

A


pH Jet-A
pH
10
9
8
7
6
10 15 20 25 30 35 40 45 50 55 60 65
Time (Days)
pH treat 1
pH treat 2
pH treat 3
pH treat 4
INDL
INDU

B


Jet-A, Average Cosine Distance
Distance
10^{0}
10^{-1}
10^{-2}
10^{-3}
Within Treat 1
Treatment 1 to Treatment 2
Treatment 1 to Treatment 3
Treatment 1 to Treatment 4
10
15
20
25
30
35
40
45
50
55
60
65
Time (Days)


Jet-A, Effect Significance
1.01
0.99
0.97
0.95
0.93
0.89
Confidence Level
Cosine vector test
Euclidean vector test
Nonmetric clustering test
10
15
20
25
30
35
40
45
50
55
60
65
Time (Days)

Nonmetric Conceptual Clustering in Ecology and Ecotoxicology

Geoffrey Matthews

Computer Science Department

Western Washington University

Robin Matthews

Huxley College of Environmental Studies

Western Washington University

Wayne Landis

Institute of Environmental Toxicology and Chemistry

Western Washington University

Send correspondence to: Geoffrey Matthews

Computer Science Department

Western Washington University, MS 9062

Bellingham, Washington, 98226, USA

matthews@cs.wwu.edu

(206) 650-3797

Running head: Nonmetric Conceptual Clustering

Abstract

Ecological studies and multispecies ecotoxicological tests are based on the examination of a variety of physical, chemical and biological data with the intent of finding patterns in their changing relationships over time. The data sets resulting from such studies are often noisy, incomplete, and difficult to envision. We have developed machine learning and visualization software to aid in the analysis, modelling, and understanding of such systems, and have applied it to the analysis of lake and stream field studies, and aquatic microcosm toxicological tests. The software is based on nonmetric conceptual clustering, which attempts to analyze the data into clusters that are strongly associated with several measured parameters. We have found in many cases that this approach is superior to classical clustering algorithms, all of which rely on an n-dimensional metric (or similarity measure). In each case, our tools not only confirmed suspected ecological patterns, but also revealed aspects of the data that were unnoticed by ecologists using conventional statistical techniques. Machine learning tools should, accordingly, become a standard part of the ecologist's armamentarium.

Introduction

Understanding ecosystems requires the solution of novel data analysis problems. Typically, dozens to hundreds of species, as well as many physical and chemical parameters, are sampled in natural and artificial systems. These parameters not only change over time, but sampling limitations necessitate acquiring only a few samples, resulting in shallow data matrices with many dimensions, but few points. The essential task of computational assistance, then, is to reduce the dimensionality and aid in the interpretation of these data sets. Nonmetric conceptual clustering was designed for these kinds of data (Matthews and Hearne, 1991). It simultaneously reduces both the complexity and the dimensionality of the set of data points. The complexity is reduced by grouping the points into clusters. The dimensionality of the data is reduced by selecting only parameters that fit well with the generated clusters. Random or noisy parameters are ignored. The ability to evaluate a model

of the data simultaneously on several different fitness criteria gives nonmetric conceptual clustering its strength.

We have applied nonmetric clustering successfully in multispecies field and laboratory studies, and in each case we have not only confirmed the presence of suspected patterns, but also discovered aspects of the data that were unnoticed by ecologists (Landis et al., 1993; Matthews et al., 1991a; Matthews et al., 1991b). In addition, these patterns were usually overlooked by conventional statistical techniques. In this sense, the software has stepped beyond the role of traditional expert systems, which merely mimic human expertise, and into the role of a machine learning system: a computer system that can learn things about the data that a human cannot. Such systems bring a new kind of power to human investigators, expertise that is beyond their own ability but which can form part of a valuable partnership.

We present here a summary of the nonmetric conceptual clustering approach, some results stemming from applications in ecology and ecotoxocology, and our attempts to extend the applicability of the nonmetric clustering paradigm to system dynamics.

Nonmetric Clustering

Nonmetric clustering is similar to conceptual clustering in that the clustering is designed, not only to fit the data, but also to create a *simple* and *conceptual* description of the data (Michalski and Stepp, 1983; Fisher and Langley, 1986). The goal of nonmetric clustering is a partition of the data into disjoint and exhaustive subsets (the clusters) such that most of the points can be described by simple conjunctive descriptions involving some of the original parameters (canonical dimensions, *i.e.* without rotation, *etc*). For example, if a large number of the points (cluster A), in dimensions x, y, and z, had "medium", "small", and "large" values, respectively, and another large number of points (cluster B), had "large", "medium", and "medium" values on these same dimensions, then the points could be described by the two concepts:

$$\text{Cluster A:} \quad \Leftrightarrow \quad (x = \text{medium}) \wedge (y = \text{small}) \wedge (z = \text{large})$$

$$\text{Cluster B:} \quad \Leftrightarrow \quad (x = \text{large}) \wedge (y = \text{medium}) \wedge (z = \text{medium})$$

If these two sets of points comprised nearly all of the original data, then the clustering would be complete. There may be other dimensions in the original data set, other than x, y, and z, but these dimensions would be regarded as irrelevant to the above clustering if x, y, and z sufficed.

To this end, the nonmetric clustering algorithm performs a (nonexhaustive) search through the space of all clusterings (partitions) of the data, and all divisions of the parameters into categories (*e.g.*, "small", "medium", and "large"), and all subsets of parameters. The search terminates when it finds a clustering, parameter subset, and categorical division, such that the fit to the data cannot be improved. Naturally, the space of partitions and divisions is too large to be searched exhaustively. Accordingly, a hill-climbing algorithm is employed, starting from a random partition and quantile divisions of the dimensions. The search is then repeated, starting from a different random initialization, to avoid local maxima. In our experience with both synthetic and real data, about ten repetitions are sufficient to avoid local maxima. The algorithm has been implemented in a computer program called Riffle, together with a graphical front end for viewing the results.

Nonmetric clustering has the following advantages over some conventional clustering methodologies:

- It works well with incomplete data, where several points may have missing values for a few dimensions.
- It works equally well with categorical, ordinal, and numeric dimensions.
- It does not require *ad hoc* modifications of the numeric dimensions, such as normalizing the variance
- It does not rely on a metric, such as the Euclidean metric, which will combine parameters by sums of squares or other mathematical methods.

- It has the ability to ignore noisy parameters, *i.e.* parameters with a large variance but random with respect to the overall pattern. Size of the variance is not taken into account since all values on all dimensions are merely regarded as small, medium, or large.

The clustering itself is informative, but Riffle actually provides the user with more than a traditional clustering algorithm. It also reports a list of the parameters that have a strong association with the clusters is also revealing. This list, which is a subset of all of the parameters, records only those that are important or significant in relation to the patterns in the data. Parameters that vary randomly are automatically be excluded from the list.

There are a number of synthetic data sets on which Riffle can outperform traditional clustering algorithms (Matthews and Hearne, 1991). However, the most amazing successes with Riffle have been in the analysis of ecological and ecotoxicological data sets, which we describe in the following sections.

Aquatic Ecology

In both lake and stream studies, Riffle has succeeded in obtaining intuitively meaningful clusters. In a one-year study of benthic macroinvertebrates in a small stream, Riffle grouped the samples exactly as a human expert would have done, one group consisting of "clean" water samples (mayflies, stoneflies, etc.), and another group consisting of "dirty" water samples (flies, oligochaetes, *etc.*) (Matthews et al., 1991a). Several rare species were found to have high association with these clusters, and thus were reported by Riffle as important to the overall pattern. But these same species had been overlooked as important indicator species because of their rarity. The samples were collected over an entire season, and included both low-density and high-density samples as the benthos matured over the summer. Standard clustering techniques were confounded by this seasonal variance and grouped the samples into "early" and "late" samples, without regard to the fine structure of the populations.

In a multi-year study of physical/chemical parameters in a large monomictic lake, Riffle accurately clustered samples according to season into summer epilimnion and

hypolimnion, as well as winter mixed water samples (Matthews et al., 1991b). In a result surprising to the investigators, it also identified a fourth class of samples. Upon reinvestigating, we noticed that this class had actually been sampled from within the metalimnion—an unforeseen accident of the experimental design. Further clustering by Riffle of the biological data showed a strong correlation with the clustering of the physical/chemical parameters. Conventional clustering algorithms were not able to identify these patterns.

Ecotoxicology

Riffle has also been successful in analyzing data from synthetic microcosms, in particular, the Standardized Aquatic Microcosm, or SAM (Taub, 1989). In the SAM, twenty-four jars of water are prepared identically with several species of algae, *Daphnia*, and other biota. The jars are divided into four treatment groups, normally a control and three increasingly toxic doses. The jars are monitored closely for two months and population counts for all species, as well as physical/chemical parameters, are recorded every few days. Nonmetric clustering by Riffle can often pick out the four treatment groups from the biological data alone.

Under controlled situations, such as the SAM, nonmetric clustering can form the basis of a confirmatory statistical test, which we have termed nonmetric clustering and association analysis (NCAA). In this case, the known treatment groups form one categorical label, and the cluster numbers form another. (Sometimes, although by no means always, the treatment groups form an ordinal, and not merely categorical variable.) The association between treatment group and cluster number forms the basis of a confirmatory statistic: under the null hypothesis, there would be no association. Any contingency table test, such as the χ^2 test, can then be used to obtain a confidence level.

Nonmetric clustering consistently reveals aspects of the SAM microcosms that are hidden from other tests. Since Riffle reduces the dimensionality of the SAM by indicating which species are important on which days of the test, it gives the practitioner a good handle on how the populations respond to the toxin. Quite often one species will be important early

in the test, of little importance during the middle period, and then important again later. We have also noticed "chaotic" trends in the evolution of the SAM. For instance, in at least two of the experiments, the treated groups diverged significantly from the control group, and then, by about the end of the first month, "recovered" to a state indistinguishable from the control group. However, during the second month, the treatment groups again diverged, in a dose-response fashion. This indicates that, during the putative recovery period, the systems were nonetheless quite different, and were able to diverge later. This is symptomatic of chaotic systems, where imperceptible differences in initial conditions can lead to radically different behavior subsequently.

Other Applications

Riffle is currently being applied to a wide variety of data analysis problems. We are currently beginning an investigation into the toxicity of refinery effluents, using measurements required by the National Pollution Discharge Elimination System (NPDES). Also, in cooperation with Dr. Anne Fairbrother of the U.S.E.P.A., Corvallis, we are applying Riffle to studies of biomarkers of toxicological impacts on mice and birds. Other researchers have applied Riffle to medical diagnosis problems.

Future Directions: Temporal Dynamics

As well as Riffle works in analyzing data, it is essentially static. Many of the effects seen in ecological data analysis are dynamic—an effect may be simply a time delay, for example. Further, oscillations, such as those in the predator-prey models, can be expected, as well as chaotic dynamics. We are beginning to apply the lessons learned from nonmetric clustering to the analysis of dynamic multivariate data. Some of our approaches are outlined below.

Discrete curvature and torsion: The path of an ecosystem through n-dimensional space over time can be viewed as a parameterized curve. Using analogies of the Frenet formulas (O'Neill, 1966, pp. 56-66), discrete analogues of the fundamental vectors, velocity, curvature, torsion *etc.*, can be defined and used to characterize the evolution of the

system.

Nonmetric clustering strain: The key idea behind nonmetric clustering strain is to measure the change in nonmetric clustering from one time slice to the next. By examining how nonmetric clusters of the points change over time, measures of the size and direction of the change can be obtained.

Conceptual shift: When performing conceptual clustering the important parameters usually change over time. Thus, not only do the points change their relationships, but the conceptual descriptions of the points can use a different vocabulary at different times. The measure of how the "best" description changes over time gives us another handle on understanding dynamic behavior.

Visualization: We are also investigating graphical visualization of the evolution of systems in n-dimensional phase space over time. The curvature, torsion, clustering shift and conceptual shift can all be visualized with interactive computer graphics. Projection pursuit and grand tour algorithms can be used to maximize the visibility of desired quantities (Asimov, 1985; Huber, 1985). Critical points, at which the behavior of the systems becomes "interesting," can then often be found by inspection.

Conclusions

Our program attempts to understand multivariate data on its own terms. To this end, we have built and applied nonmetric clustering and visualization tools that reduce the dimensionality and complexity of multispecies systems to a manageable size. Other attempts have been made to understand ecosystems in terms of multivariate response, but the responses were usually measured using n-dimensional metrics (Johnson, 1988; Kersting, 1988). We have seen repeatedly that metric approaches suffer from a large number of drawbacks when dealing with ecological data. The approach recommended here is free from any metric (or similarity measure) and its problems.

Recently, the U.S. Environmental Protection Agency has instituted a policy that calls for the cancellation of multispecies toxicity tests because data analysis has proven too difficult or inconclusive (Fisher, 1992). We believe that the problem is not with the multispecies tests, which are carefully designed to be more realistic than classic, single-species tests, but rather with the poor quality of the data analysis tools that are applied to the results of these tests. So far as we know, we are the only group in the United States applying the methodologies of machine learning to multivariate ecological and ecotoxicological studies, and we are seeing results that greatly enhance our understanding of the systems and their dynamics. Interest in our techniques at national toxicological conferences is always high, and we are convinced that the machine learning paradigm will revolutionize ecology and ecotoxicology in the near future.

References

Asimov, D. (1985). The grand tour. *SIAM Journal of Scientific and Statistical Computing*, 6:128–143.

Fisher, D. and Langley, P. (1986). Conceptual clustering and its relation to numerical taxonomy. In Gale, W. A., editor, *Artificial Intelligence and Statistics*, pages 77–116. Addison Wesley.

Fisher, L. J. (1992). Decisions of the ecological, fate, and effects task force. *U.S. E.P.A. Memo, October 29.*

Huber, P. J. (1985). Projection pursuit. *Annals of Statistics*, 13:435–475.

Johnson, A. R. (1988). Evaluating ecosystem response to toxicant stress: a state space approach. In Adams, W. J., Chapman, G. A., and Landis, W. G., editors, *Aquatic Toxicology and Hazard Assessment: 10th Volume, ASTM STP 971*, pages 275–285. American Society for Testing and Materials, Philadelphia.

Kersting, K. (1988). Normalized ecosysterm strain in micro-ecosystems using different sets of state variables. *Verhandlungen der Internationalen Vereinigung fuer Theoretische und Angwandte Limnologie*, 23:1641–1646.

Landis, W. G., Matthews, R. A., Markiewicz, A. J., and Matthews, G. B. (1993). Multivariate analysis of the impacts of the turbine fuel jp-4 in a microcosm test with implications for the evaluation of ecosystem dynamics and risk assessment. *Ecotoxicology.*

Matthews, G. and Hearne, J. (1991). Clustering without a metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(2):175–184.

Matthews, G. B., Matthews, R. A., and Hachmöller, B. (1991a). Mathematical analysis of temporal and spatial trends in the benthic macroinvertebrate communities of a small stream. *Canadian Journal of Fisheries and Aquatic Sciences*, 48(11):2184–2190.

Matthews, R. A., Matthews, G. B., and Ehinger, W. J. (1991b). Classification and ordination of limnological data: a comparison of analytical tools. *Ecological Modelling*, 53:167–187.

Michalski, R. S. and Stepp, R. E. (1983). Learning from observation: Conceptual clustering. *Machine Learning, An Artificial Intelligence Approach*, pages 331–363.

O'Neill, B. (1966). *Elementary Differential Geometry.* Academic Press, New York.

Taub, F. B. (1989). Standardized aquatic microcosms. *Environm. Sci. Technol.*, 23:1064–1066.

Multivariate Analysis of the Impacts of the Turbine Fuel JP-4 in a Microcosm Toxicity Test with Implications for the Evaluation of Ecosystem Dynamics and Risk Assessment

Wayne G. Landis*, Robin A. Matthews, and April J. Markiewicz
Institute of Environmental Toxicology and Chemistry, Huxley College of Environmental Studies
Western Washington University, Bellingham, WA 98225.

Geoffrey B. Matthews
Computer Science Department
Western Washington University, Bellingham, WA 98225.

* To whom correspondence should be addressed

Running Title: Multivariate Analyses of JP-4 Toxicity

Corresponding Author: Wayne G. Landis, Institute of Environmental Toxicology and Chemistry, Huxley College of Environmental Studies, Western Washington University, Bellingham, WA 98225.
Phone: 206-650-6136, Fax: 206-650-7284, Internet: landis@henson.cc.wwu.edu

Abstract: Turbine fuels are often the only aviation fuel available in most of the world. Turbine fuels consist of numerous constituents with varying water solubilities, volatilities and toxicities. This study investigates the toxicity of the water soluble fraction (WSF) of JP-4 using the Standard Aquatic Microcosm (SAM). Multivariate analysis of the complex data, including the relatively new method of nonmetric clustering, was used and compared to more traditional analyses. Particular emphasis is placed on ecosystem dynamics in multivariate space.

The WSF is prepared by vigorously mixing the fuel and the SAM microcosm media in a separatory funnel. The water phase, which contains the water-soluble fraction of JP-4 is then collected.

The SAM experiment was conducted using concentrations of 0.0, 1, 5 and 15 percent WSF. The WSF is added on day 7 of the experiments by removing 450 mL from each microcosm including the controls, then adding the appropriate amount of toxicant solution and finally bringing the final volume to 3L with microcosm media. Analysis of the WSF was performed by purge and trap gas chromatography (Figure 2). The organic constituents of the WSF were not recoverable from the water column within several days of the addition of the toxicant. However, the impact of the WSF on the microcosm was apparent. In the highest initial concentration treatment group an algal bloom ensued, generated by the apparent toxicity of the WSF of JP-4 to the daphnids. As the daphnid populations recovered the algal populations decreased to control values. Multivariate methods, clearly demonstrated this initial impact along with an additional oscillation separating the 4 treatment groups in the latter segment of the experiment. Apparent recovery may be an artifact of the projections used to describe the multivariate data. The variables that were most important in distinguishing the four groups shifted during the course of the 63 day experiment. Even this simple microcosm exhibited a variety of dynamics, with implications for biomonitoring schemes and ecological risk assessments.

Key Words-Jet fuel, microcosm, multivariate statistics, nonmetric clustering, risk assessment

Introduction

As this is written, the United States Environmental Protection Agency has suspended the requirement for conducting ecosystem level studies for pesticide registration (Fisher, 1992). Although many factors contributed to the action, apparently the field and pond mesocosm tests that were conducted did not contribute to the evaluation of risk of pesticides in a timely and cost effective manner.

Over the last 15 years a variety of multispecies toxicity tests have been developed with the hope that in doing so, the increased complexity of the test would result in more realistic, community-level responses to the toxicant. However, the addition of more than one species, and the generally longer time periods associated with these multispecies tests, also result in much more complex data sets. Distinguishing toxicant effects from other community-level changes has become one of the most critical obstacles to the interpretation of multispecies data sets.

Multispecies toxicity tests are usually referred to as microcosms or mesocosms, although a clear definition of the size or complexity to distinguish these terms has not been put forth. Multispecies toxicity tests range from approximately 1 L (e.g., mixed flask cultures) to thousands of liters, as in the case of the pond mesocosms used in pesticide registration testing. The number of species and origin of those taxa can vary widely. In the Standardized Aquatic Microcosm (SAM) developed by Taub and colleagues (Taub, 1969, 1976; Taub and Crow, 1978; Crow and Taub, 1979; Taub *et al.*, 1980; Kindig, 1983; Taub, 1987; Taub, *et al.*, 1988; Taub, 1988, 1989; Conquest and Taub, 1989) the physical, chemical, and biological components are defined as to species, media and substrate (see Table 1 and Figure 1). In other systems colonization by the importation of sediment or by repeated inoculation from a natural source is used to establish the model system. Larger systems often use a combination of means to start and maintain a multispecies, interactive community.

One of the major difficulties in the evaluation of multispecies toxicity tests has been the difficulty in the analysis of the large data set on a level consistent with the goals of the toxicity test. Typically, the goals of the toxicity test are:

- to detect changes in the population dynamics of the individual taxa that would not be apparent in single species tests; and,
- to detect community-level differences that are correlated with treatment groups thereby representing a deviation from the control group.

A number of methods have been developed to attempt to satisfy the goals of multispecies toxicity testing. Analysis of variance (ANOVA) is the classical method to examine single variable differences from the control group. However, because multispecies toxicity tests generally run for weeks or even months, there are problems with using conventional ANOVA. These include the increasing likelihood of introducing a Type II error (accepting a false null-hypothesis), temporal dependence of the variables, and

the difficulty or graphically representing the data set. Conquest and Taub (1989) developed a method to overcome some of the problems by using intervals of non-significant difference (IND). This method corrects for the likelihood of Type II errors and produces intervals that are easily graphed to ease examination. The method is routinely used to examine data from SAM toxicity tests, and it is applicable to other multivariate toxicity tests. The major drawback is the examination of a single variable at a time over the course of the experiment. While this addresses the first goal in multispecies toxicity testing, listed above, it ignores the second. In many instances, community-level responses are not as straightforward as the classical predator/prey or nutrient limitation dynamics usually picked as examples of single-species responses that represent complex interactions.

Multivariate methods have proved promising as a method of incorporating all of the dimensions of an ecosystem. One of the first methods used in toxicity testing was the calculation of ecosystem strain developed by Kersting (1984, 1985, 1988) for a relatively simple (three species) microcosm. This method has the advantage of using all of the measured parameters of an ecosystem to look for treatment-related differences. At about the same time, Johnson (1988a, 1988b) developed a multivariate algorithm using the n-dimensional coordinates of a multivariate data set and the distances between these coordinates as a measure of divergence between treatment groups. Both of these methods have the advantage of examining the ecosystem as a whole rather than by single variables, and can track such processes as succession, recovery and the deviation of a system due to an anthropogenic input.

However, a major disadvantage of both these methods, and of many conventional multivariate methods, is that all of the data are often incorporated without regard to the units of measurement or the appropriateness of including all variables in the analysis. It can be difficulty to combine variables such as pH, with units ranging from 0-14, with the numbers of bacterial cells per ml, where low numbers are in the 10^6 range, to say nothing of the conceptual difficulties of adding pH units to counts. Similarly, random variables (i.e., variables with not treatment-related response) indiscriminately incorporated into the analysis may contribute so much noise that they overshadow variables that do show treatment-related effects.

Ideally, a multivariate statistical test used for evaluating complex data sets will have the following characteristics:

- It will not combine counts from dissimilar taxa by means of sums of squares, or other *ad hoc* mathematical techniques, as in the Euclidean and cosine distance measures.
- It will not require transformations of the data, such as normalizing the variance.
- It will work without modification on incomplete data sets.
- It will work without further assumptions on different data types (e.g., species counts or presence/absence data).
- Significance of a taxon to the analysis will not be dependent on the absolute size of its

count, so that taxa having a small total variance, such as rare taxa, can compete in importance with common taxa, and taxa with a large, random variance will not automatically be selected, to the exclusion of others.

- It will provide an integral measure of "how good" the analysis is, i.e. whether the data set differs from a random collection of points.
- It will, in some cases, identify a subset of the taxa that serve as reliable indicators of the physical environment.

Recently developed for the analysis of ecological data, nonmetric clustering is a multivariate derivative of artificial intelligence research that satisfies all these criteria, and has the potential of circumventing many of the problems of conventional multivariate analysis.

In this paper, we use ANOVA and intervals of non-significant difference, and three multivariate techniques to search for meaningful patterns in the data set from a SAM toxicity test using Jet-A turbine fuel. The multivariate techniques include two conventional tests based on the ratio of multivariate metric distances (Euclidean distance and cosine of the vector distance), and one relatively new program, RIFFLE, which employs nonmetric clustering and association analysis (Matthews and Hearne, 1991). All three of the multivariate techniques have proven useful in analyzing complex ecological data sets (Matthews *et al.*, 1991; Matthews *et al.*, 1991). Of the three, only nonmetric clustering meets all of the criteria listed above (Matthews and Matthews, 1991). The major disadvantage of the RIFFLE program is that, in order to find a clustering of the data points with the desirable qualities listed above, a massive search through thousands of potential clustering candidates is made before settling on the "right" one. Even after this search, there is no guarantee that RIFFLE finds an optimal clustering. However, in our experience, RIFFLE does find an excellent clustering in reasonable time.

Jet fuels or perhaps more accurately, turbine fuels, are one of the primary fuels for internal combustion engines worldwide and certainly are the most widely available aviation fuel. Over the last 15 years virtually all of the commercial airline operations and charter operations have converted to a turbine engine because of the inherent low operating cost of the power plant, its reliability, and in part to the availability of fuel even in underdeveloped areas. In the U.S. military there has been a progressive replacement of conventional piston engine vehicles with turbine equivalents. Standardization on a single type of turbine fuel to relieve logistical demands is also underway. Given the overwhelming predominance of turbine fuel, a fuel spill or accidental release of aviation fuel will likely be one of the prevalent turbine fuels: Jet-A, used for commercial and general aviation; JP-4, the standard fuel of the U.S. Air Force and Army Aviation; and JP-5, the naval equivalent of JP-4. JP-8 is a new fuel proposed as the standard for all military vehicles using turbine engines.

Along with the environmental considerations, turbine fuels also offer advantages as model complex toxicants for toxicological research. Because of their use as aviation fuel, turbine fuels are produced to

stringent specifications designed to ensure the safety of flight. Therefore, the overall general properties of these materials are tightly controlled. In addition, standard archived samples of the military fuels are maintained for toxicological studies at Wright Patterson, AFB. Jet fuels also tend to be less explosive and also less volatile than gasoline, making the materials easier and safer to use. Like all petroleum products, however, the exact identity of the constituents varies according to the original crude and the refining process.

This paper reports the effects of low concentration of the water soluble fraction of JP-4 on the community incorporated in the SAM. The effects of the WSF on the microcosm communities were subtle. An early increase in algal density was apparent in the treatment groups containing the highest concentrations of the WSF and was matched by a decrease in daphnid populations. Multivariate analysis proved to be more powerful and efficient in highlighting important variables and processes than ANOVA. The variables that were most important were those distinguishing where treatment-related effects shifted during the course of the experiment. The multivariate analysis also detected oscillations in the similarity of the control and dosed groups that were not apparent using conventional univariate tests. The oscillations may be due to the inherent perturbations in community dynamics and interactions, or the effects upon the segments of the community not directly measured, the bacterial detritivores. We also discuss the implications of this research with regards to the use of indices and the conduct of environmental risk assessments.

Methods and Materials

Reagents

All chemicals used in the culture of the organisms and in the formulation of the microcosm media were reagent grade or as specified by the ASTM method.

JP-4 was supplied by the U. S. Air Force Toxicology Laboratory at Wright Patterson, AFB Ohio.

Water Soluble Fraction

The water soluble fraction of JP-4 was prepared in glassware washed in nonphosphate soap, rinsed, then soaked in 2N HCl for at least one hour, rinsed ten times with distilled water, dried and finally autoclaved for 30 minutes. Microcosm medium, T82MV, acted as the diluent for the water fraction of the WSF.

Twenty five mL of JP-4 is added to the two liter separatory funnel, and is agitated as follows: [1] Shake separatory funnel for five minutes, releasing built up pressure as necessary, [2] allow funnel contents to remain undisturbed for 15 minutes, [3] shake contents for five minutes, allow to stand 15 minutes, [4] continue same pattern for a total time of 1 hour, and finally [5] allow separatory funnel contents to remain undisturbed for eight hours. At the end of this procedure the mixture was allowed to stand overnight. The next day all but 100 mL of T82MV/water soluble fraction of jet fuel mixture from the

separatory funnel (leaving the lighter, insoluble fuel mixture in the flask) was drained into a cleaned, sterile 1 liter amber glass bottle and capped with a Teflon-lined screw cap. The WSF was used within twenty-four hours or stored at 4°C for no longer than forty-eight hours before use as toxicant mixture.

Gas Chromatography of WSF

This protocol utilizes a Tekmar LSC 2000 Purge and Trap (P&T) concentrator system in tandem with a Hewlett Packard 5890A Gas Chromatograph with a Flame Ionization Detector (FID) (ASTM D3710, 1988; ASTM D2887, 1988; Westendorf, 1986). Instrument blanks and deionized distilled water blanks are used to verify the P&T and GC columns cleanliness prior to analysis of samples. A five mL sample is injected into a five milliliter sparger, purged with pre-purified nitrogen gas for eleven minutes and dry purged for four minutes. Volatile hydrocarbons, purged from the sample and collected on the Tenax/Silica Gel column, are desorbed at 180°C directly onto the gas chromatograph SPB-5, 30m x 0.53 mm ID 1.5µm film, fused silica capillary column. The column, at 35°C, is held at that temperature for two minutes, increased to 225°C at 12°C/min and held at that temperature for five minutes. A Spectra-Physics 4290 Integrator records the FID signal output of the volatile hydrocarbons that have been separated and eluted from the column by molecular weight.

Identification and quantification of GC fractions

Qualitative identification of some components in the water soluble fraction (WSF) of the JP-4 fuel, used as the toxicant in the microcosm test, were determined using a Simulated Distillation (SIMDIS) Calibration Mixture. The ASTM Method D3710 Qualitative Calibration Mixture is the standard test method for determining the Boiling Range Distribution of Gasoline and Gasoline Fractions by Gas Chromatography. This mixture was used as a calibration standard to determine the retention times for each known component in the mixture against which unknown components, in the WSF of the Jet fuel mixture, were compared and identified.

Quantitative estimates of some components of the WSF were made by comparing sample chromatographs to certified n-paraffin and n-naphtha chromatograph standards, prepared and analyzed under the same P&T/GC conditions.

Algal Toxicity Tests

In order to estimate the relative toxicities of the JP-4 mixture and to set the concentrations for the microcosm a series of short-term toxicity tests were performed (ASTM E 1218, 1991). Algal growth inhibition tests were performed using *Ankistrodesmus falcatus* and *Selenastrum capricornutum* strains identical to those used in the SAM toxicity tests.

Test algae were grown in a semi-flow through culture apparatus on the microcosm media T82MV and taken during log phase growth for inoculation into the test flasks. Two hundred and fifty (250) ml

Erlenmeyer flasks were used as test chambers, with serial dilutions of the water soluble fraction at concentrations of 0.0, 6.25, 12.5, 25, 50 and 100 percent then placed in the flasks. The test organisms were added at a concentration of approximately 3.0×10^4 cells/mL. Total volume was 100 mL with two replicates of controls and the test concentrations used. Test mixtures will be incubated at 20.0°C ± 1.0°C with a 12:12 hour light/dark cycle. Using a Newbauer Counting Chamber, cell densities were determined every 24 hrs for the 96 hr duration of the test.

The cell numbers are then plotted against the concentrations. If possible, a least square regression line was drawn and the IC_{50} (the concentration at which algal growth is inhibited to 50% of the control) determined. ANOVA is then run on the replicates to determine if any of the groups are significantly different.

SAM Protocol

The 64-day SAM protocol previously has been described (ASTM E 1366-91, 1991). Table 1 describes the organisms, conditions and modifications of ASTM E1366-91 for this particular experiment. Briefly, the microcosms were prepared by the introduction of ten algal, four invertebrate, and one bacterial species into 3 L of sterile defined medium. Test containers were 4 L glass jars. An autoclaved sediment consisting of 200 g silica sand and 0.5 g of ground chitin is autoclaved in the experimental jar immersed in a water bath to a point above the sand and chitin level during sterilization. This procedure helps prevent breakage of the jars and subsequent loss of replicates.

Numbers of organisms, dissolved oxygen (DO) and pH were determined twice weekly. Room temperature was 20°C ± 2°. Illumination was 79.2 μEm^{-2} sec^{-1} PhAR with a range of 78.6-80.4 and a 16/8 day/night cycle.

Two major modifications were made to the SAM protocol. The first was the means of toxicant delivery. Test material was added on day 7 by stirring each microcosm, removing 450 mL from each container and then adding appropriate amounts of the WSF to produce concentrations of 0, 1, 5 and 15 percent WSF. After toxicant addition the final volume was adjusted to 3L. No attempt was made to filter and retain the organisms withdrawn during the removal of the 450 mL prior to toxicant addition. All graphs and statistical analysis start with the first sampling day, day 11.

The second modification was the substitution of *Tetrahymena thermophila* BIV for the hypotrichous ciliate used in past experiments. The hypotrichous ciliate was becoming increasingly difficult to culture, very likely due to the age of the clone. *T. thermophila* has routinely been used in biochemical research and in detoxification studies of organophosphates (Landis *et al.* 1985, 1987, 1991). Using SAM controls, constructed prior to this experiment, it was demonstrated that the *T. thermophila* populations were able to exist within the system. *T. thermophila* are maintained sterilly in a 3 percent proteous peptone distilled water media at 20°C with routine biweekly transfers to perpetuate the stocks. The results presented below demonstrate the suitability of the Tetrahymena for inclusion in the protocol.

Data Analysis

All data were recorded onto standard computer entry forms and checked for accuracy. The data was then keyed into the SAMS data analysis program and checked for accuracy. Parameters calculated included the concentrations of each of the species, DO, DO gain and loss, net photosynthesis/respiration ratio (P/R), pH, algal species diversity, algal biovolume, and biovolume of available algae. The statistical significance of these parameters compared to the controls was also computed for each sampling day using the Interval of Non-significant Difference (IND) plots developed by Conquest. Note that algal biovolume, algal species diversity and available algae are all derived variables based on the algal counts. The net photosynthesis/respiration ratio is not derived using ^{14}C methods but by comparing oxygen concentrations before lights on, at the end of the photosynthetic period, and then at the next morning, as specified in the standard protocol. Photosynthesis/respiration ratio was the variable used during the analysis to incorporate these measurements.

The multivariate methods used in the analysis include cosine and vector distances and nonmetric clustering. All of these methods have been previously described (Matthews *et al.*, 1991; Landis *et al.*, 1993, Landis *et al.*, 1993) and are reviewed in Appendix A. Table 2 lists the variables used in the clustering process.

Results

Algal Toxicity Results. The WSF of JP-4 was not particularly toxic when used as a *percentage (v/v)* of the total culture media. As determined by graphical analysis, since 100 percent inhibition was not achieved, the IC_{50} for Ankistrodesmus was 57 percent WSF and for Selenastrum 95 percent WSF.

Persistence of the JP-4 WSF. Seven compounds, benzene, 2,4 dimthylpentane, ethylbenzene, 2-methylpentane, 2-methylpropane, o-xylene and toluene, were tracked using GC analysis during the course of the SAM experiment. Figure 2 is an area graph that presents both the concentrations of the individual components along with the totals of these seven materials in microcosms of Treatment 4. As can be readily seen, 504 hrs after dosing, the relative concentrations of these materials have rapidly disappeared. After week three, only 2-methylpentane and 2-methylpropane are detectable. Since only the 2-methylpropane is present 672 hours after dosing, this material may be the final biodegradative product of the absorbed fraction of the WSF, and is being investigated in more detail.

Patterns in Algal Communities. The largest increase in algal population density occurred in treatment 4 (Figure 3). The peak density is approximately twice that of the control replicates at day 21. After the initial bloom in treatment 4, no particular dose-related pattern is discernible. Lyngbya makes up a substantial portion of the algal community in each treatment group, which is historically unusual. Algal species diversity also generally declines in each of the treatment groups, but in a general sense not related to dose.

Daphnid Populations. Each of the treatment groups exhibited similar dynamics (Figure 4). None of

the groups were statistically different from the control groups using conventional analysis of variance approaches. Minor perturbations in the timing of the peaks may have occurred, but by day 50 the means of each group were very similar.

Ostracod Populations. At the end of the experiment, the average population density in the control treatments is approximately twice that of treatment 4, the highest toxicant concentration (Figure 5). Population density in the two treatment groups with the highest toxicant concentrations, decline below the no dose treatment and the lowest treatment densities. This pattern is apparent graphically from day 53 onward. Conventional analysis such as the IND plot does not pick any date as significantly different from the control. The probability of the order remaining consistent on five consecutive dates by chance alone and assuming independence is small $((0.25)^4)^4$).

Philodina and Tetrahymena Populations. Tetrahymena survived in each of the treatment groups until near the end of the experiment (Figure 6a). No specific dose related pattern was apparent although a two sampling period bloom (days 25 and 27) was apparent for Treatment 2. Unfortunately the error in sampling and the inherent asynchrony in Protistan reproduction prevented the result from being detectable using conventional methods. Philodina did not appear in appreciable numbers until after day 25 in any of the treatments. Day 53 showed a dramatic increase in treatments 3 and 4 followed by a decline, so that by day 60 all treatments were similar. Although suggestive, the results are not significant; the large overlap of the standard deviation apparent (Figure 6b). The difficulty in sampling rapidly growing and declining populations in asynchronous growth is apparent. Although trends may be suggested, conventional analysis does not detect a significant effect.

pH and Photosynthesis/Respiration ratio. Treatment 4 pH did exhibit a statistically significant difference from the other treatments during the period of the algal bloom during the first ten days after dosing (Figure 7a). On day 49 a deviation from the control in a dose response manner was detected. However with the multiple comparisons being made it is difficult to attribute such an event to the treatment. At the end of the experiment all of the groups resembled reference treatment.

The photosynthesis/respiration ratio (Figure 7b) did not exhibit statistically significant differences during the course of this experiment.

Multivariate results. The significance levels for the three multivariate tests performed for each sampling day are graphed in Figure 8. All tests agree, that a significant difference between treatment groups was observed through day 25. Nonmetric clustering demonstrated fluctuations in this significance from day 25 until 40, and from 40 until the end of the experiment. The cosine vector and Euclidean vector methods were statistically significant until after day 53.

In Figure 9, the average cosine distances within the reference group and between the reference group and each of the three treatment groups are plotted on a log scale. The initial, strong effect, from day 11 to day 25, is easily seen as a large distance from the reference treatment 1 (no dose) and treatment 4 (highest dose). The period from day 25 to 30 reflect another more subtle oscillation that is

statistically significant using cosine vector and Euclidean vector clustering. From day 35 to day 46 the distances from treatment 1 to the other treatments are similar to the within treatment 1 distances and the nonmetric clustering does not detect a significant difference. A third period of separation from the control that is statistically significant using the distance measures, from day 46 to 53, is seen for the JP-4 SAM.

Also of interest are the variables that best described the clusters and the stability of the importance of the variables during the course of the experiment. Table 3 lists the variables determined to be important in defining the clusters of importance for each sampling day as determined by nonmetric clustering. In general, the number of variables that were important was larger during the start of the test and lower at the end. In addition, a great deal of variability in rankings is apparent during the course of the SAM. The number of sampling dates when a variable was deemed important in cluster formation is listed in Table 4. Chlorella and S. Daphnia were ranked 8 out of the 16 sampling dates with Ankistrodesmus ranked 6 out of 16, being ranked in 12 out of the 16 sampling dates. The distribution of ranks was rather even although variables such as Tetrahymena and Ulothrix did not appear.

The timing of each variable gaining importance in the determination of clusters was also interesting. Ostracods and Philodina were important after day 32 of the experiment, as were small Daphnia. Chlorella was selected as a significant variable throughout the course of the experiment.

Discussion

The examination of individual parameters provided only a limited and somewhat distorted view of the dynamic responses of the SAM system to JP-4. The univariate data did show that there was some significant responses to the toxicant as determined by the chemistry. Biological data, taken individually, did not demonstrate a coherent and unified picture of the response of the biota to JP-4. The biological responses that were most evident were of only dramatic impacts, such as the increase in the algal populations due to the inhibitory effect of the JP-4 upon the grazer populations. Axiomatically, an inhibition of the predominant grazer in the early stages of the microcosm, the Daphnia, is going to result in an algal bloom. These types of responses do not provide a depth of understanding of the function and structure of the artificial ecosystem. In contrast to the biological data, pH did demonstrate some statistically significant differences using the IND methodology that hinted at an early major impact in treatment 4 and a later divergence. It is likely that pH is measuring an alteration in the metabolism of the system and therefore a change in the functionality, but without structural differences it is difficult to attribute the functional differences to structural alterations.

The multivariate analyses of the structural data revealed patterns not observed using the univariate analysis of the biotic data. Three oscillations from the non dosed treatment 1 could be observed that were statistically significant. Two of these oscillations correspond well to the divergences seen in the pH analysis. However in the divergences seen between days 25-30 and 50-55 (Figure 9), suggestions of a dose-response can be seen that are not apparent in the pH data. It is important that these oscillations

were observed after the demise of the original WSF mixture, no doubt lost to volatilization or biotransformation and degradation by the biota.

A similar set of results have been obtained for a related toxicant, Jet-A (Landis *et al.*, 1993). In a virtually identical experiment, univariate methods were able to demonstrate alteration in the grazer (daphnid)-algal dynamics and in two functional measures, pH and P/R ratio. Subsequent departures of the dosed treatments from the non dosed treatments were not observed using the biotic measures. However, the functional measures, pH and P/R, both demonstrated an additional divergence for one sampling date in the latter half of the microcosm experiment. However, the univariate analysis does not corroborate these results and they may have been dismissed as chance occurrences without the multivariate analyses.

The multivariate analyses depicted at least two statistically significant oscillations using all three measurement techniques. As with the Jet-A, the original WSF mixture had rapidly decreased in concentration during the first few week after dosing.

A detailed comparison of the dynamics of the two SAM experiments is currently underway to compare similarities and differences in the multivariate space of the impacts of the two mixtures. However, changes in the structural composition the systems did occur repeatedly during the course of the experiments even in these relatively simple systems. These oscillations point to effects not readily observed or predicted by single species systems. The repeated divergence of the dosed systems from the reference systems can be accounted for in two ways:

- It may reflect the functioning of the community in terms of parameters not directly sampled by the SAM protocol.

- It may be a persistent fluctuation in the community structure initiated by the initial stress, but is only periodically visible, as if it were an incompletely dampened nonlinear oscillation in the systems' inherent dynamics.

Examination of individual parameters provides only a limited, and somewhat distorted view of the SAM microcosm response to the WSF of each fuel. The univariate data analysis did indeed show that there were some significant responses to the toxicant by individual taxa and chemistry; however, the responses were scattered over time, and did not present a logical, coherent pattern. Furthermore, the individual responses detected were typified by wild swings in the population density of a taxon over time.

If you kill or restrict the reproduction of most of the Daphnia, the next microcosm response is likely an algal bloom. This result could is easily have been predicted by the short term toxicity tests and was expected. However, recent modeling efforts by Taub *et al.* (submitted) suggest that the dynamics of these interactions and the resulting magnitudes of the algal blooms are highly dependent upon the timing

of the toxic insult. Measuring these types of gross responses to the toxicant do not provide much more insight into impact of the toxicant in the ecosystem than do the short-term single-species tests. The absolute magnitude of the disturbance and the period of recovery can be obtained from the microcosm experiment, in the sense of a classical predator prey interaction. However, the multivariate analysis reveals a more interesting dynamic.

The multivariate patterns suggest a much more complex pattern of multiple divergences and convergences in the similarities between treatment groups. Much as an ecosystem could be expected to display the rise and fall of species assemblages, the SAM microcosms appear to indicate that the first divergence is only the beginning of a series of responses.

Using nonmetric clustering, we can list the variables that were the most important for separating the treatment group clusters for each day that measurements were collected (Table 3). The list of variables suggests that the first divergence, which occurred from about day 11 through day 25, results from predator/prey interactions between primary producers (algae) and first order consumers (daphnia). This divergence should be characterized by the following properties:

- The divergence will be fast, because the algae and daphnia populations are introduced into the microcosm after being cultured in optimal laboratory conditions and then placed into cultures with high available nutrient concentrations. Predation, or the lack of predation, or other limiting factors will cause rapid changes in the algal and herbivore populations.

- The divergence will be short-lived, because the populations are unstable in the nutrient rich early successional microcosm. There will be a tendency for the microcosms to drift away from the early "treatment" effect into a more typical community based on both algae and detritus as the food source for the secondary consumers. Initially, this drift may mask treatment effects and be interpreted as recovery of the system.

The first divergence is the only type of response that is normally searched for in microcosm tests using conventional statistics. This response is typical of many reported SAM experiments (Taub *et al.*, 1988; Taub, 1988; Haley *et al.*, 1988; Landis *et al.*, 1989).

The second and third divergences occurred from between days 25-30 and 50-55 . During this time, Daphnia and some of the algal taxa were often still important in the cluster development; however, other secondary consumers (Ostracods and Philodina) entered the list. The second divergence may represent the long-term effects of the initial toxicant on a more successionally mature community that is fueled by both algal productivity and detritus. If so, the resulting divergences should have the following characteristics:

- It should be strongly influenced by detritus quality. Detritus is conditioned by bacteria and fungi, which are highly sensitive to toxins but are unmeasured in the microcosm. Also, detritus that has passed through the gut of a consumer (e.g., consumed algae) is different than detritus that originates directly from dead algae (unconsumed). Therefore, the quality of the detritus may be highly affected by the treatment, but none of the factors influencing the effects will be measured directly.

- Secondary consumers of detritus and bacteria are no less affected by the quality of their food source than algal consumers, so the treatment-related alterations of the quality of detritus and bacteria will cause differences in the secondary consumer populations.

Therefore, the series of divergences following the initial algal-daphnid interaction may still represent a direct response to the initial treatment effects, but because it occurs late in the microcosm experiment, it is easily misinterpreted as noisy or the effects of a degradation product. An inclusion of measures of detritus quality and microbial metabolism may answer these questions and such studies are currently being incorporated into our series of microcosm experiments.

Invoking unseen properties of an ecosystem or other mechanistic explanations may not be needed to explain the occurrence of oscillations and divergences from a non-dosed reference system. An alternative and complimentary explanation is available that perhaps describes the dynamics of multispecies systems at a more fundamental level.

First, the apparent recovery or movement of the dosed systems towards the reference or treatment 1 case may be an artifact of our measurement systems that allow the n-dimensional data to be represented in a two dimensional system. In an n-dimensional sense, the systems may be moving in opposite directions and simply pass by similar coordinates during certain time intervals. Positions be similar but the n-dimensional vectors describing the movements of the systems can be very different.

The apparent recoveries and divergences may also be artifacts of our attempt to chose the best means of collapsing and representing n-dimensional data into a two or three dimensional representation. In order to represent such data, it is necessary to project n-dimensional data into three or less dimensions. As information is lost when the shadow of a cube is projected upon a two dimensional screen, a similar loss of information can occur in our attempt to represent n-dimensional data. The possible illusion of recovery based on this type of projection is diagramatically represented in Figure 10. In Figure 10a the dosed and the reference systems appear to converge, i. e. recovery has occurred. However, this may be an illusion created by the perspective chosen to describe and measure the system. Figure 10b is the same system but viewed from the "top". When a new point of view is taken, divergence of the systems occurs throughout the observed time period. As the various groups separate, the divergence may be seen as a separate event. In fact, this separation is a continuation of the dynamics

initiated earlier upon one aspect of the community. Eventually, the illusion of recovery may simply be the divergence of the replicates within each treatment group becoming large enough, with enough inherent variation, so that even the multivariate analysis can not distinguish treatment group similarities. Not every divergence from the control treatment may have a causal effect related to it in time; differentiating these events from those due to degradation products or other perturbations will be challenging.

Not only may system recovery be an illusion but there are strong theoretical reasons that seem to indicate that recovery to a reference system may be impossible or at least unlikely. In fact, systems that differ only marginally in their initial conditions and at levels probably impossible to measure, are likely to diverge in unpredictable manners. May and Oster (1978) in a particularly seminal paper investigated the likelihood that many of the dynamics seen in ecosystems, generally attributed to chance or stochastic events, are in fact deterministic. In fact simple deterministic models of populations can give rise to complicated behaviors. Using equations resembling those used in population biology, bifurcations occur resulting with several distinct outcomes. Eventually, given the proper parameters, the system appears chaotic in nature although the underlying mechanisms are completely deterministic. Obviously, biological systems have limits, extinction being perhaps the most obvious and best recorded. Another ramification is that the noise in ecosystems and in sampling may not be the result of a stochastic process but the result of underlying deterministic, but chaotic relationships.

These principals also apply to spatial distributions of populations as recently reported by Hassell *et al.* (1991). In a study using host-parasite interactions as the model, a variety of spatial patterns were developed using the Nicholson-Bailey model. Host-parasite interactions demonstrated patterns ranging from static 'crystal lattice' patterns, spiral waves, chaotic variation or extinction with the appropriate variation of only three parameters within the same set of equations. The deterministically determined patterns could be extremely complex and not distinguishable from stochastic environmental changes.

Given the perhaps chaotic nature of populations it may not be possible to predict accurately species presence, population interactions, or structural and functional attributes. Kratz *et al.* (1987) examined the spatial and temporal variability in zooplankton data from a series of five lakes in North America. Much of the analysis was based on limnological data collected by Brige and Juday from 1925 to 1942. Copepods and cladocera, except *Bosmina*, exhibited larger variability between lakes than between years in the same lake. Some taxa showed consistent patterns among the study lakes. They concluded that the controlling factors for these taxa operated uniformly in the each of the study sites. However, in regards to the depth of maximal abundance for calanoid copepods and Bosmina, the data obtained from one lake had little predictive power for application to other lakes. Part of this uncertainty was attributed to the intrinsic rate of increase of the invertebrates with variability increasing with a corresponding increase in r_{max}. A high r_{max} should enable the populations to accurately track changes in the environment. Katz *et al.* suggest that these taxa be used to track changes in the environment. Unfortunately, in the context of environmental toxicology, the inability to use one lake to predict the non-dosed population dynamics of

these organisms in another, reduces the sensitivity of methods that use comparisons of two systems as measures of anthropogenic impacts.

A better strategy may be to let the data and a clustering protocol identify the important parameters in determining the dynamics of and impacts to ecological systems. This approach has been recently suggested independently by Dickson *et al.* (1992) and Matthews and Matthews (Matthews *et al.*, 1991; Matthews and Matthews, 1991). This approach is in direct contrast to the more usual means of assessing anthropogenic impacts. One classical approach is to use the presence or absence of so called indicator species. This assumes that the tolerance to a variety of toxicants is known and that chaotic or stochastic influences are minimized. A second approach is to use hypothesis testing to differentiate metrics from the systems in question. This second approach assumes that the investigators know *a priori* the important parameters. Given that, at least in our relatively simple SAM systems, the important parameters in differentiating non-dosed from dosed systems changes from sampling period to sampling period, this assumption can not be made. Classification approaches such as nonmetric clustering or the canonical correlation methodology developed by Dickson *et al.* eliminates these assumptions.

The results presented in this report combined with the others cited above and the implications of chaotic dynamics suggest that reliance upon any one variable or an index of variables may be an operational convenience that may provide a misleading representation of pollutant effects and the associated risks. The use of indices such as diversity and the Index of Biological Integrity have the effect of collapsing the dimensions of the descriptive hypervolume in a relatively arbitrary fashion. Indices, since they are composited variables, are not true endpoints. The collapse of the dimensions that are composited tends to eliminate crucial information, such as the inherent variability, and its importance in describing these variables. The mere presence or absence and the frequency of these events can be analyzed using techniques such as nonmetric clustering that preserve the nature of the dataset. A useful function was certainly served by the application of indices, but the new methods of data compilation, analysis and representation derived from the Artificial Intelligence tradition can now replace these approaches and illuminate the underlying structure and dynamic nature of ecological systems. In the next 18 months RISC based computers will make these approaches widely available at the desktop level.

The implications are important. Currently, only small sections of ecosystems are monitored or a heavy reliance is placed upon, so-called, indicator species. These data suggest that, to do so is dangerous, potentially producing misleading interpretations and resulting in costly error in management and regulatory judgments. Much larger toxicological test systems are currently analyzed using conventional statistical methods on the limit of acceptable statistical power. Interpretation of the results has proven to be difficult.

The dynamics observed in our experiments and in the research discussed above should make obvious that a metaphor such as ecosystem health is inappropriate and misleading. In a recent critical evaluation, Suter (1993) dismissed ecosystem health as a misrepresentation of ecological science.

Ecosystems are not organisms with the patterns of homeostasis determined by a central genetic core. Since ecosystems are not organismal in nature, health is a property that can not describe the state of such a system. The urge to represent such a state as health has lead to the compilation of variables with different metrics, characteristics and casual relationships. Suter suggests a better alternative would be to evaluate the array of ecosystem processes of interest, with an underlying understanding that the fundamental nature of these systems are quite different than those of organisms.

One of the ongoing debates in environmental toxicology has been the suitability of the extrapolation and realism of the various multispecies toxicity tests that have been developed over the last 15 years. One of the major criticisms of small scale systems is that the low diversity of the system is not representative of natural systems in dynamic complexity (Sugiura, 1992). Given the above discussion and the conclusions derived from it much of this debate may have been misdirected. The small scale systems used in our study have been demonstrated to express complex dynamics. Kersting and Van Wungaarden (1992) found that even the three compartment microecosystem, as developed by Kersting (1984, 1985, 1988), expresses indirect effects as measured by pH changes after dosing with chloropyrifos. Since even full scale systems can not serve as reliable predictors of the dynamics of other full scale systems, it is impossible to suggest that any artificially created system can provide a generic representation of any full scale system. Debate should probably revert to more productive areas such as improvements in culture, sampling and measurement techniques or other characteristics of these systems. A more worthwhile goal is probably the understanding of the scaling factors, in a full n-dimensional representation, that should enable the accurate representation of specific ecosystem characteristics. Certain aspects of a community may be included in one system to answer specific questions that in another system would be entirely inappropriate. If questions as to detritus quality are important then the system should include that particular component. In other words, the system should attempt to answer the particular scientific question.

Several questions are now the goals of future research. The dynamics of the loss of jet fuels from the SAM systems is currently being investigated in greater depth. Additional data should indicate the persistence of the constituents and help aid in the determination of initial toxicity, including further information from literature searchs or using quantitative structure activity relationship models. Additional testing of related materials is being conducted. Finally, questions as to the effects of size and community structure abound. The SAM system is relatively simple. Data sets incorporating more diverse species assemblages and of varying sizes are being investigated for comparison.

Conclusions

1. Effects are seen in the microcosm study that can only in part be attributed to the differential toxicity. At least three oscillations are distinguishable from the reference system related to treatment.

2. Multivariate analysis is crucial in observing effects with typically noisy datasets and points to the dynamic nature of the variables important in distinguishing the four treatment groups.

3. Two general hypotheses are proposed to account for the observed dynamics of the system. The oscillations may be result of structural and functional components not measured, such as detrital processing and quality. The second and not exclusive hypothesis is that the oscillations are due to the inherent chaotic nature of ecosystems and may propagate in an unpredictable fashion over time.

4. The implications of these results is that reliance upon indices that condense data or upon indicator species may be misleading in determining effects of stressors upon biological communities. A strategy providing better resolution in determining ecosystem impacts may be the sampling of a broader set of variables, accepting the variability inherent in sampling, since it may be impossible due to the nature of the system to predict relevant measurements. If it is inherently impossible to predict the relevant parameters, only an examination of a compendium of data from the system is likely to reliably measure effects.

5. If multiple undampened oscillations and chaotic dynamics characterize ecosystems then concepts such as ecosystem health and ecosystem recovery should be eliminated or redefined. Chaotic systems are unlikely to exhibit characteristics that correspond to the health at the organismal level. Similarly, recovery of a system to a preexisting state may be impossible or highly unlikely.

Appendix A. Multivariate Techniques-Nonmetric Clustering

In the research described above, three multivariate significance tests were used. Two of them were based on the ratio of multivariate metric distances within treatment groups vs. between treatment groups. One of these is calculated using Euclidean distance and the other with cosine of vectors distance (Good, 1982; Smith *et al.*, 1990). The third test used nonmetric clustering and association analysis (Matthews and Matthews, 1990). In the microcosm tests there were four treatment groups with six replicates, giving a total of 24. This example is used to illustrate the applications in the derivations that follow.

Treating a sample on a given day as a vector of values, $\bar{x} = \langle x_1, \ldots x_{17} \rangle$, with one value for each of the measured biotic parameters, allows multivariate distance functions to be computed. Euclidean distance between two sample points $\bar{x}$ and $\bar{y}$ is computed as

$$\sqrt{\sum_i (x_i - y_i)^2}$$

The cosine of the vector distance between the points $\bar{x}$ and $\bar{y}$ is computed as

$$1-\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum y_i^2}}$$

Subtracting the cosine from one yields a distance measure, rather than a similarity measure, with the measure increasing as the points get farther from each other.

The within-between ratio test used a complete matrix of point-to-point distance (either Euclidean or cosine) values. For each sampling date, one sample point $\bar{x}$ was obtained from each of six replicates in the four treatment groups, giving a 24 x 24 matrix of distances. After the distances were computed, the ratio of the average within group metric (*W*) to the average between group metric (*B*) was computed (*W/B*). If the points in a given treatment group are closer to each other, on average, than they are to points in a different treatment group, then this ratio will be small. The significance of the ratio is estimated with an approximate randomization test (Noreen, 1989). This test is based on the fact that, under the null hypothesis, assignment of points to treatment groups is random, the treatment having no effect. The test, accordingly, randomly assigns each of the replicate points to groups, and recomputes the *W/B* ratio, a large number of times (500 in our tests). If the null hypothesis is false, this randomly derived ratio will (probably) be larger than the *W/B* ratio obtained from the actual treatment groups. By taking a large number of random reassignments, a valid estimate of the probability under the null hypothesis is obtained as $(n+1)/(500+1)$, where n is the number of times a ratio less than or equal to the actual ratio was obtained (Noreen, 1989).

In the clustering association test, the data are first clustered independently of the treatment group, using nonmetric clustering and the computer program RIFFLE (Matthews and Hearn, 1991). Because the RIFFLE analysis is naive to treatment group, the clusters may, or may not correspond to treatment effects. To evaluate whether the clusters were related to treatment groups, whenever the clustering procedure produced four clusters for the sample points, the association between clusters and treatment groups was measured in a 4 x 4 contingency table, each point in treatment group i and cluster j being counted as a point in frequency cell ij. Significance of the association in the table was then measured with Pearson's X^2 test, defined as

$$\chi^2 = \sum_{ij} \frac{\left(N_{ij} - n_{ij}\right)^2}{n_{ij}}$$

where N_{ij} is the actual cell count and n_{ij} is the expected cell frequency, obtained from the row and column marginal totals N_{+j} and N_{i+} as

$$n_{ij} = \frac{N_{+j}N_{i+}}{N}$$

where $N = 24$ is the total cell count (Press *et al.*, 1990), and a standard procedure for computing the significance (probability) of χ^2 taken from Press (1990).

Table 1. Summary of Test Conditions for Conducting SAM JP-4.

Organisms	
Organisms per chamber:	Algae (added on Day 0 at initial concentration of 10^3 cells for each algae species): *Anabaena cylindrica*, *Ankistrodesmus* sp., *Chlamydomonas reinhardi* 90, *Chlorella vulgaris*, *Lyngbya* sp., *Scenedesmus obliquus*, *Selenastrum capricornutum*, *Stigeoclonium* sp., and *Ulothrix* sp. Animals (added on Day 4 at the initial numbers indicated in parentheses): *Daphnia magna* (16/microcosm), *Cypridopsis* sp. (ostracod) (6/microcosm), *Tetrahymena thermophila* [protozoa] (0.1/mL), and *Philodina* sp. (rotifer) (0.03/mL)
Experimental design	
Test vessel type and size:	One-gallon (3.8 L) glass jars16.0 cm wide at the shoulder, 25 cm tall with 10.6 cm openings
Medium volume:	3000 mL added to each container
Number of replicates x concentrations:	6x4
Reinoculation:	Once per week add one drop (circa 0.05 mL) to each microcosm from a mix of the ten species = 5×10^2 cells of each alga added per microcosm
Addition of test materials:	Test material added day 7 by removing 450 mL from each container and then adding appropriate amounts of the WSF to produce concentrations of 0, 1, 5 and 15 percent WSF. After toxicant addition the final volume was adjusted to 3L.
Sampling frequency:	2 times each week
Test duration:	63 days
Physical and chemical parameters	
Temperature:	20 to 25°C
Light intensity:	80 µE m^{-2} photosynthetically active radiation s^{-1} (850 to 1000 fc)
Photoperiod:	12 h light/12 h dark
Medium:	Medium T82MV
Sediment:	Composed of silica sand (200 g), ground, crude chitin (0.5), and cellulose powder (0.5 g) added to each container
Measurements:	Algal, invertebrate and protozoa counts, pH, dissolved oxygen, optical density. Parameters calculated included the concentrations of each of the species, DO, DO gain and loss, net photosynthesis/respiration ratio (P/R), pH, algal species diversity, daphnid fecundity, algal biovolume, and biovolume of available algae.

Table 2. Biotic parameters used in the multivariate statistical tests. Biotic variables such as diversity, available biovolume, and total algal biovolume are not used since they are derived from and therefore not independent of the variables listed above.

- Anabaena
- *Ankistrodesmus*
- Chlamydomonas
- Chlorella
- Daphnia
  - Ephipia
  - Small Daphnia
  - Medium Daphnia
  - Large Daphnia
- Tetrahymena
- Lyngbya
- Miscellaneous sp.
- Ostracod (*Cyprinotus*)
- Philodina (Rotifer)
- Scenedesmus
- Selanastrum
- *Stigeoclonium*
- Ulothrix

Table 3. Important variables as determined by nonmetric clustering ranked according to contribution for each sampling day. Some variables such as Ankistrodesmus were important in determining group clusters in the first half of the experiment. Some of the variables such as **Ostracod** and Philodina were more important in the latter stages of the experiment. Note that the order of importance of even the more common contributors often changed from sampling day to sampling day, with no one variable being consistently ranked, *Chlorella* and S. Daphnia being the closest.

Day	Important Variables in Determining Clusters in Rank Order
11	Selanastrum, M. Daphnia, *Chlorella*, Ankistrodesmus
14	Selenastrum, S. Daphnia, M. Daphnia-Ankistrodesmus[1], L. Daphnia-Stigeoclonium
18	Scenedesmus, Selanstrum, Ankistrodesmus, S. Daphnia, *Chlorella*, L. Daphnia
21	Scenedesmus, Ankistrodesmus, Chlamydomonas
25	*Chlorella*, S. Daphnia
28	*Chlorella*, Ankistrodesmus-Lyngbya, Philodina
32	**Ostracod**
35	**Ostracod**, Philodina, Scenedesmus
39	Scenedesmus, S. Daphnia
42	Lyngbya, S. Daphnia, Philodina, Ankistrodesmus
46	M. Daphnia
49	Scenedesmus, Chlorella, Philodina
53	*Chlorella*, Philodina
56	M. Daphnia-S. Daphnia
60	S. Daphnia, **Ostracod**, Lyngbya
63	*Chlorella*, S. Daphnia, M. Daphnia, Lyngbya

[1] Hyphen between variables denotes equal rank

Table 4. Variable According to Success in Determining Clusters as Defined by Nonmetric Clustering. Variables such as Ankistrodesmus and the Daphnia classes were important in the course of this study. However, reliance on any particular organism or a small combination would have poorly described the dynamics of the system.

Variable	Ranked
Chlorella	8
S. Daphnia	8
Ankistrodesmus	6
Scenedesmus	5
Philodina	5
M. Daphnia	4
Lyngbya	4
L. Daphnia	3
Ostracod	3
Selenastrum	3

Figures

Figure 1. Timeline for the Standardized Aquatic Microcosm JP-4 Experiment. Each step of this 63 day protocol is choreographed according to ASTM E 1366-91. The modifications to the protocol are the elimination of *Nitchia*, *Hyalella azteca* , modification of the method for toxicant delivery and the substitution of *T. thermophila* BIV for the hypotrichous ciliate.

Figure 2. Purge and Trap Gas Chromatography Results for the WSF of JP-4. A substantial reduction in the number and concentration of the WSF constituents is apparent two weeks after dosing in Treatment 4. At the end of the SAM experiment the fractions are at relatively low concentrations.

Figure 3. Patterns in Algal Communities. The largest increase in algal population density occurred in treatment 4 (Figure 3d). The peak density is approximately twice that of the control replicates at day 21. After the initial bloom in treatment 4 no particular dose-related pattern is discernible.

Figure 4. Daphnid Population Dynamics. Each of the treatment groups exhibited similar dynamics (Figure 4). None of the groups were statistically different from the control groups using conventional analysis of variance and IND approaches. Minor perturbations in the timing of the peaks may have occurred, but by day 49 the means of each group a very similar.

Figure 5. Ostracod Population Dynamics. The average population density in the control treatments is approximately twice that of Treatment 4, the highest concentration. In between, the populations densities are ranked in a dose response manner. Although suggestive and not readily apparent in the other biological data, the apparent dose response falls within the IND plot surrounding the control. The bars are standard deviations for the means of each sampling day. An IND is approximately 2.5 times the standard deviation.

Figure 6. Tetrahymena and Philodina Population Dynamics. The population dynamics of the Philodina suggest a treatment effect towards the end of the experiment. As with the ostracods the sampling error is too large to distinguish such an effect using conventional univariate techniques. The bars are standard deviations for the means of each sampling day. An IND is approximately 2.5 times the standard deviation.

Figure 7. pH. Treatment 4 pH did exhibit a statistically significant difference from the reference treatment during the period of the algal bloom during the first ten days after dosing (INDL = IND upper limit, INDV = IND upper limit). On day 49 an additional deviation from the control in a dose response manner was

detected.

Figure 8. Significance levels of the three multivariate statistical tests for each sampling day. Note that there are two periods, early and late ones, where the clustering into treatment groups is significant at the 95 percent confidence level or above.

Figure 9. Cosine distance from the control group to each of the treatments for each sampling day. Note that large differences are apparent early in the SAM. During the middle part of the 63 day experiment the distances between the replicates of Treatment 1, the control group, is as large as the distances to the treatment groups. However, later in the experiment the distances from the dosed microcosms to the control again increase followed by another apparent convergence.

Figure 10. Diagrammatic representation of ecosystem movements in ecosystem space. In Figure 10a the dosed and the reference systems appear to converge, i. e. recovery has occurred. However, this may be an illusion of the variables chosen to describe the system. Figure 10b is the same system but viewed from the "top". When a new point of view is taken, divergence of the systems occurs throughout the observed time period.

Acknowledgements

We would like to thank the support of the Institute of Environmental Toxicology and Chemistry support staff, notably L. Keel, R. Sahakian and N. Shough. L. Holmquist and T. Litwiller contributed to the editing and production of this document. This research is supported by USAFOSR Grant No. AFOSR-91-0291 DEF.

References

ASTM D3710 (1988) Standard Test Method for Boiling Range Distribution of Gasoline and Gasoline Fractions by Gas Chromatography, 1988 Annual Book of ASTM Standards, Vol. 5.03, pp 78-88. American Society for Testing and Materials, Philadelphia.

ASTM D2887 (1988) Standard Test Method for Boiling Range Distribution of Petroleum Fractions by Gas Chromatography, 1988 Annual book of ASTM Standards, Vol. 5.02, pp 506-513. American Society for Testing and Materials, Philadelphia.

ASTM E 1218 (1991) Conducting Static 96-h Toxicity Tests with Microalgae. Annual book of ASTM Standards, Vol. 11.04, pp 845-856. American Society for Testing and Materials, Philadelphia.

ASTM E 1366-91 (1991) Standard Practice for the standardized aquatic microcosm: fresh water, Vol 11.04. pp 1017-1051. American Society for Testing and Materials, Philadelphia.

Conquest, L.L. and Taub, F.B. (1989) Repeatability and reproducibility of the Standard Aquatic Microcosm: Statistical properties. In *Aquatic Toxicology and Hazard Assessment: 12th Volume, ASTM STP 1027* (Cowgill, U.M. and Williams, L.R., eds) American Society for Testing and Materials, Philadelphia, PA, pp. 159-177.

Crow , M.E. and Taub, F.B. (1979) Designing a microcosm bioassay to detect ecosystem level effects. *Intern. J. Environmental Studies.* 141-147.

Dickson, K.L., Waller, W.T., Kennedy, J.H. and Ammann, L.P. (1992) Assessing the relationship between ambient toxicity and instream biological response. *Env. Tox. Chem.* **11**, 1307-1322.

Fienberg, S.E. (1985) *The Analysis of Cross-Classified Categorical Data.* MIT Press, Cambridge, MA.

Fisher, L. (1992) Memorandum: Decisions on the Ecological, Fate and Effects Task Force. Office of Pesticides and Toxic Substances, U. S. Environmental Protection Agency.

Good, I.J. (1982) An index of separateness of clusters and a permutation test for its significance. *J. Statist. Comp. Simul.* **15**, 81-84.

Haley, M.V., Johnson, D.W. and Landis, W.G. (1988) The aquatic toxicity of brass dust. In *Aquatic Toxicology and Environmental Fate: Tenth Volume ASTM STP -971* (Adams, W., Chapman, G. and Landis, W.G., eds) American Society for Testing and Materials, Philadelphia. pp 468-479.

Hassell, M.P.H, Comins, N. and May, R.M. (1991) Spatial structure and chaos in insect population dynamics. *Nature* **353**, 255-258.

Johnson, A.R. (1988a) Evaluating ecosystem response to toxicant stress: a state space approach. In *Aquatic Toxicology and Hazard Assessment: 10th Volume, ASTM STP* 971 (Adams, W.J., Chapman, G.A. and Landis, W.G., eds) American Society for Testing and Materials, Philadelphia, pp. 275-285.

Johnson, A.R. (1988b) Diagnostic variables as predictors of ecological risk. *Environmental Management* **12**, 515-523.

Katz, T.K., Frost, T.M. and Magnuson, J.J. (1987) Inferences from spatial and temporal variability in ecosystems: Long-term zooplankton data from lakes. *Amer. Nat.* **129**, 830-846.

Kersting, K. (1984) Development and use of an aquatic micro-ecosystem as a test system for toxic substances. Properties of an aquatic micro-ecosystem IV. *Int. Revue ges. Hydrobiol.* **69**, 567-607.

Kersting, K. (1985) Properties of an aquatic micro-ecosystem V. Ten years of observations of the prototype. *Verh. Internat. Verein. Limnol.* **22**, 3040-3045.

Kersting, K. (1988) Normalized ecosystem strain in micro-ecosystems using different sets of state variables. *Verh. Internat. Verein. Limnol.* **23**, 1641-1646.

Kersting , K., and van Wungaarden, R. (1992) Effects of Chlorpyifos on a microecosystem. *Env. Tox, Chem.* **11**, 365-372.

Kindig, A.C., Loveday, L.C. and Taub, F.B. (1983) Differential sensitivity of new versus mature synthetic microcosms to streptomycin sulfate treatment. In *Aquatic Toxicology and Hazard Assessment: Sixth Symposium, ASTM 802.* (Bishop, W.E., Cardwell, R.D. and Heidolph, B.B., eds) American Society for Testing and Materials, Philadelphia, pp. 192-203.

Landis, W.G., Chester, N.A., Haley, M.V., Johnson, D.W., Muse, Jr., W.T. and Tauber, R.M. (1989) The utility of the standard aquatic microcosm as a standard method for ecotoxicological evaluation. In *Aquatic Toxicology and Environmental Fate: Eleventh Volume ASTM STP -1007* (Suter, G. and Adams, M., eds) American Society for Testing and Materials, Philadelphia pp 353-367.

Landis, W.G., Haley, M.V. and Chester, N.A. (1993) The use of the standardized aquatic microcosm in the evaluation of degradative bacteria in reducing impacts to aquatic ecosystems. In *Environmental Toxicology and Risk Assessment: First Volume, ASTM STP -1179* (Landis, W.G., Hughes, J. and Lewis, M., eds) *In press,* American Society for Testing and Materials, Philadelphia, in press.

Landis, W.G., Matthews, R.A., Markiewicz, A.J., Shough, N.J. and Matthews, G.B. (1993) Multivariate analysis of the impacts of turbine fuel using a standard aquatic microcosm toxicity test. *J. Env. Sci.* In Press.

Matthews, G.B. and Hearne, J. (1991) Clustering without a metric. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, **13**, 175-184.

Matthews, G.B. and Matthews, R.A. (1991) A model for describing community change. In *Pesticides in Natural Systems: How Can Their Effects Be Monitored? Proceeding of the Conference*, Environmental Research Laboratory/ORD, Corvallis, OR, EPA 9109/9-91/011.

Matthews, G.B., Matthews, R.A. and Hachmoller, B. (1991) Mathematical analysis of temporal and spatial trends in the benthic macroinvertebrate communities of a small stream. *Canadian Journal of Fisheries and Aquatic Sciences*. **48**, 2184-2190.

Matthews, R.A., Matthews, G.B. and Ehinger, W.. .1991) Classification and ordination of limnological data: a comparison of analytical tools. *ological Modeling*. **53**, 167-187.

May, R.M. and Oster, G.F. (1978) Bifurcations and dynamical complexity in simple ecological models. *Amer. Nat.* **110**, 573-599.

Noreen, E.W. (1989) *Computer Intensive Methods for Testing Hypotheses*. Wiley-Interscience, New York, NY.

Press, W.H., Flannery, B.P., Teukolsky, A.A. and Vetterline, W.T. (1990) *Numerical Recipes in C, the Art of Scientific Computing*. Cambridge University Press, New York, NY.

Smith, E.P., Pontasch, K.W. and Cairns, Jr., J. (1990) Community similarity and the analysis of multispecies environmental data: a unified statistical approach. *Water Res.* **24**, 507-514.

Sugiura, K. (1992) A multispecies laboratory microcosm for screening ecotoxicological impacts of chemicals. *Env. Tox. Chem.* **11**, 1217-1226.

Suter, G. (1993) A critique of ecosystem health: Concepts and indices. *Environ Tox. Chem.* in press.

Taub, F.B. (1969) Gnotobiotic models of freshwater communities. *Verh Internat. Verein. Limnol.* **17**, 485-496.

Taub , F.B. (1976) Demonstration of pollution effects in aquatic microcosms. *Intern J. Environmental Studies*. **10**, 23-33.

Taub, F.B. (1988) Standardized aquatic microcosm - development and testing. *Aquatic Ecotoxicology* II.

Taub, F.B. (1989) Standardized aquatic microcosms. *Environm. Sci. Technol.* 23, 1064-1066.

Taub, F.B. and Crow, M.E. (1978) Loss of a critical species in a model (laboratory) ecosystem. *Verh. Internat. Verein. Limnol.* 1270-1276.

Taub, F.B. , Crow, M.E. and Hartmann, H.J. (1980) Responses of aquatic microcosms to acute mortality. *Microcosms in Ecological Research.* Giesy, J.P. Jr., Technical Information Center, U. S. Department of Energy. Washington, D.C., 513-535.

Taub, F.B., Kindig, A.C. and Conquest, L.L. (1987) Interlaboratory testing of a standardized aquatic microcosm. In *Aquatic Toxicology and Hazard Assessment: 10th Volume, ASTM STP 971* (Adams, W.J., Chapman, G.A. and Landis, W.G., eds) American Society for Testing and Materials, Philadelphia, PA, pp. 385-405.

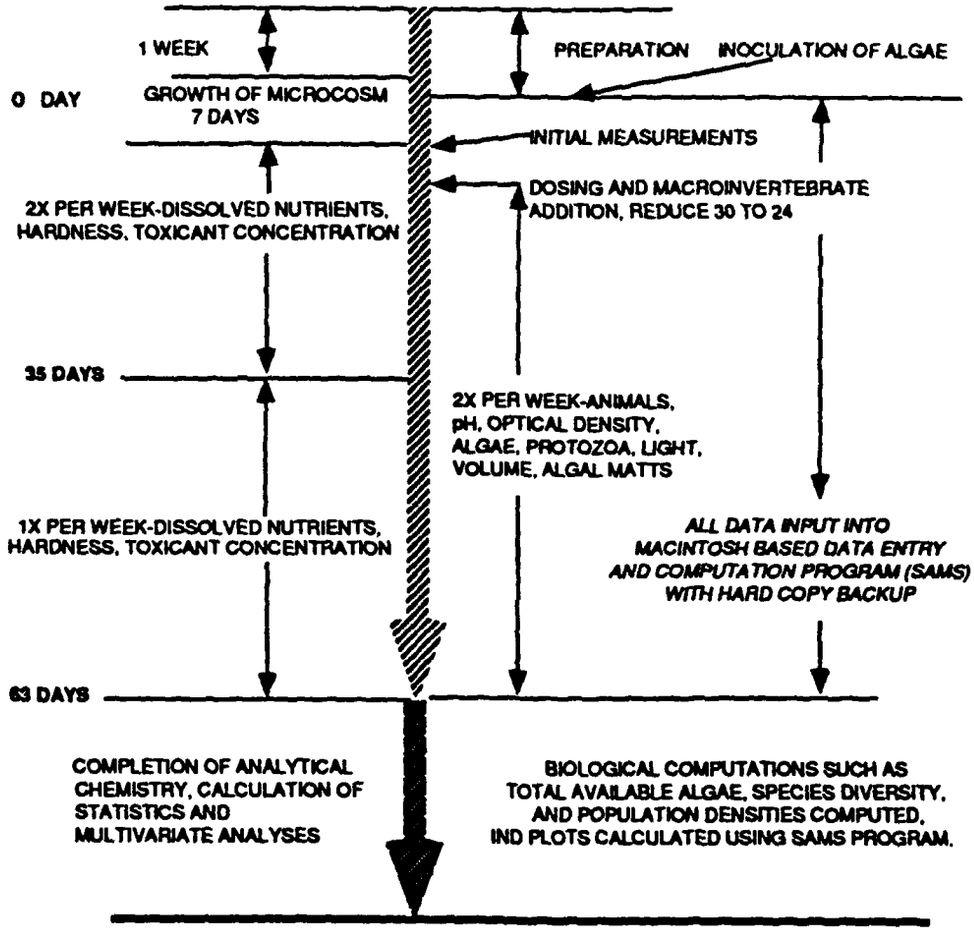
Taub, F.B., Kindig, A.C., Conquest, L.L. and Meador, J.P. (1988) Results of the interlaboratory testing of the Standardized Aquatic Microcosm protocol. In *Aquatic Toxicology and Hazard Assessment: Eleventh Symposium, ASTM* (Suter, G. and Lewis, M., eds) American Society for Testing and Materials, Philadelphia, PA.

Taub , F.B. and Read, P.L. (1983): Standardized Aquatic Microcosm Protocol: Final Report Contract No. 223-80-2352, Vol II. Food and Drug Administration. Washington, D.C.

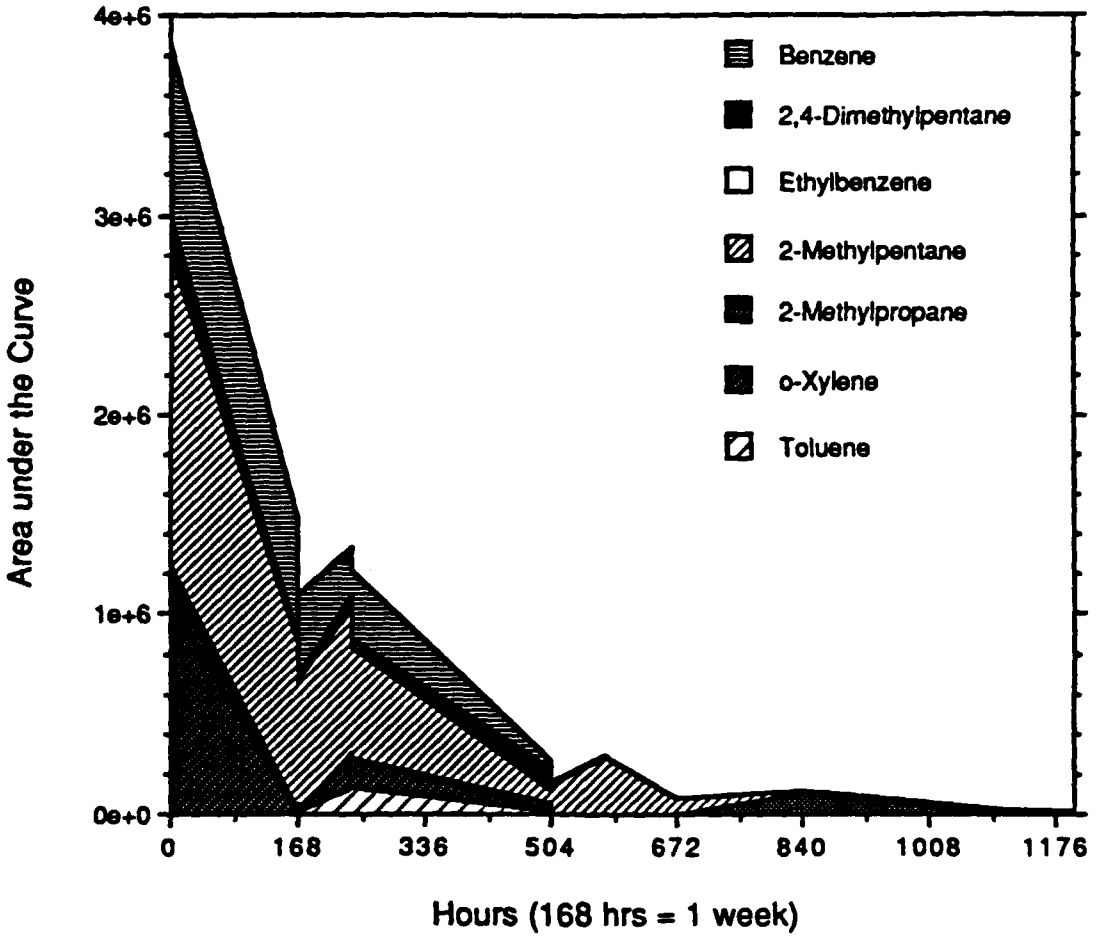
Taub, F.B., Rose, K.A., Swartzman, G.L. and Taub, J.H. (submitted) Translating population toxicity to community effects. *Env. Tox. Chem.*

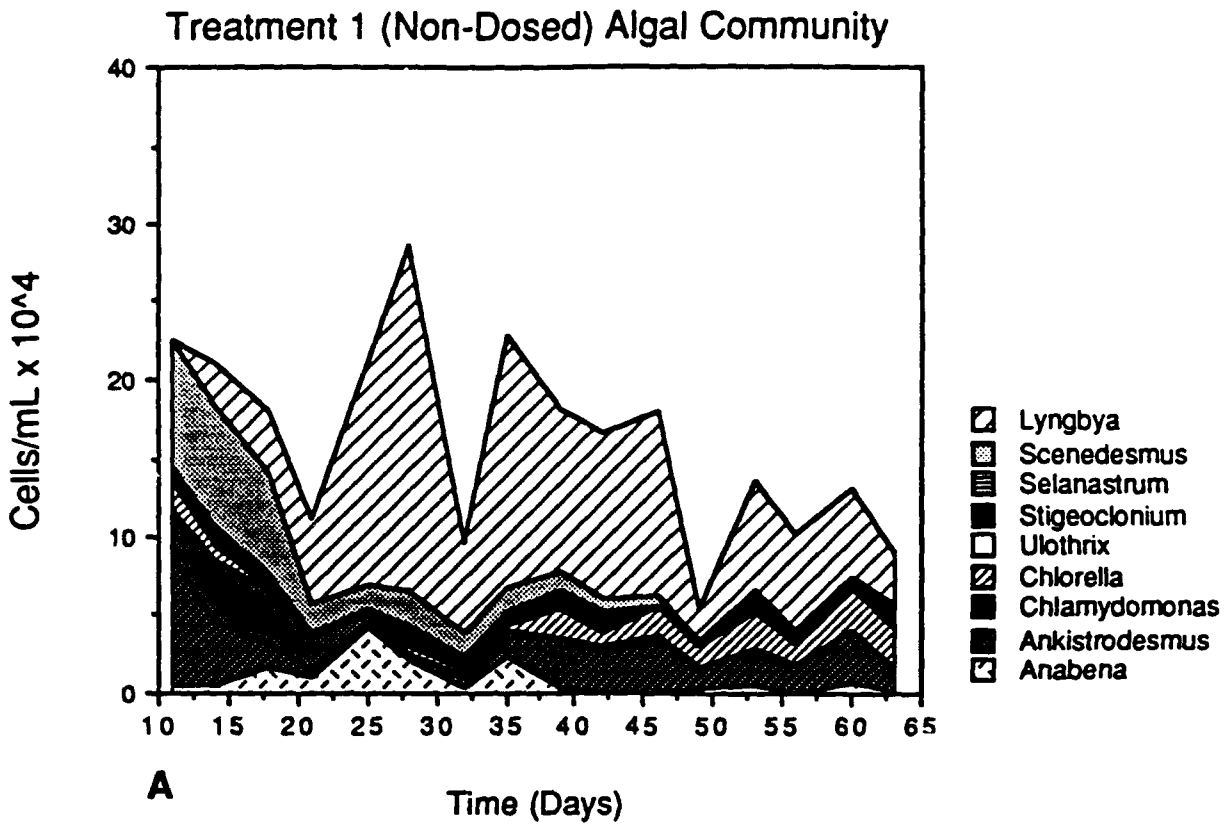
Westendorf, R.G. (1986) Performance aspects of volatile organics analysis by purge and trap capillary column gas chromatography with flame ionization detectors. Tekmar Technical Papers, Tekmar Company, Cincinnati, Ohio.

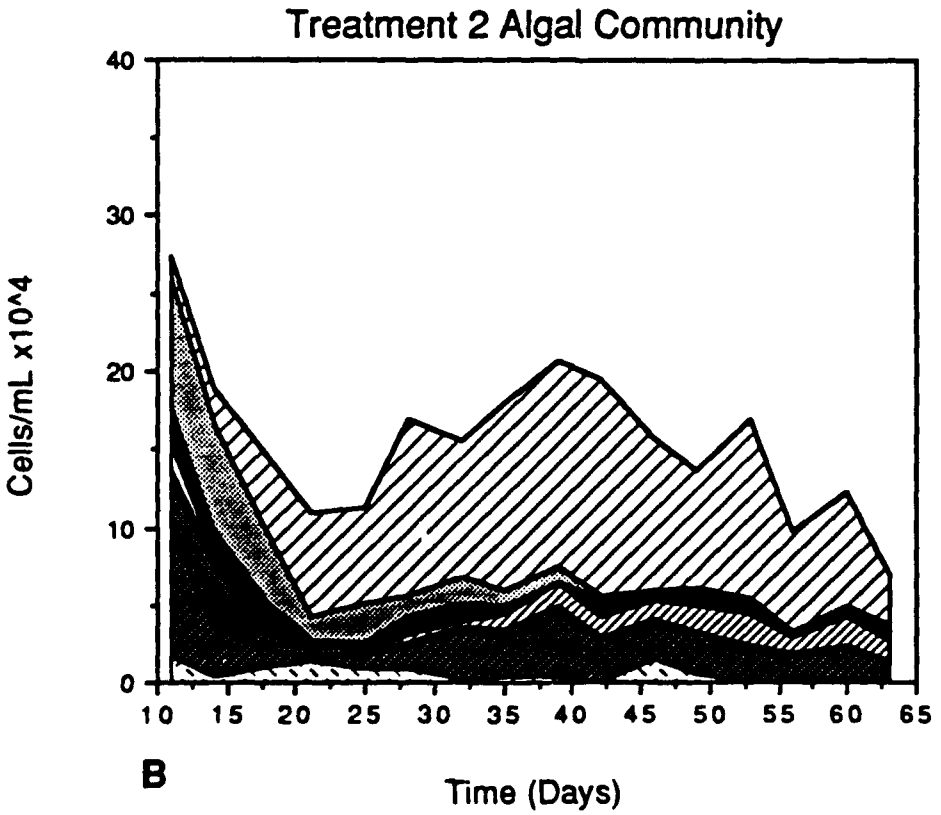
Time Line-Standardized Aquatic Microcosm

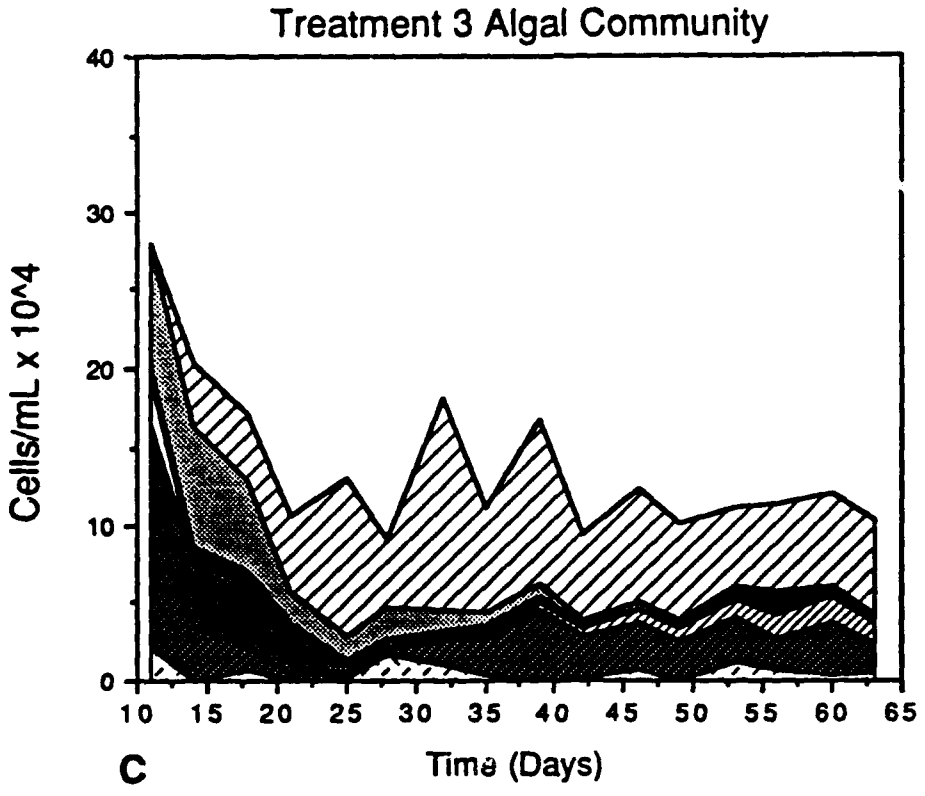


JP-4 WSF Degradation Area Graph

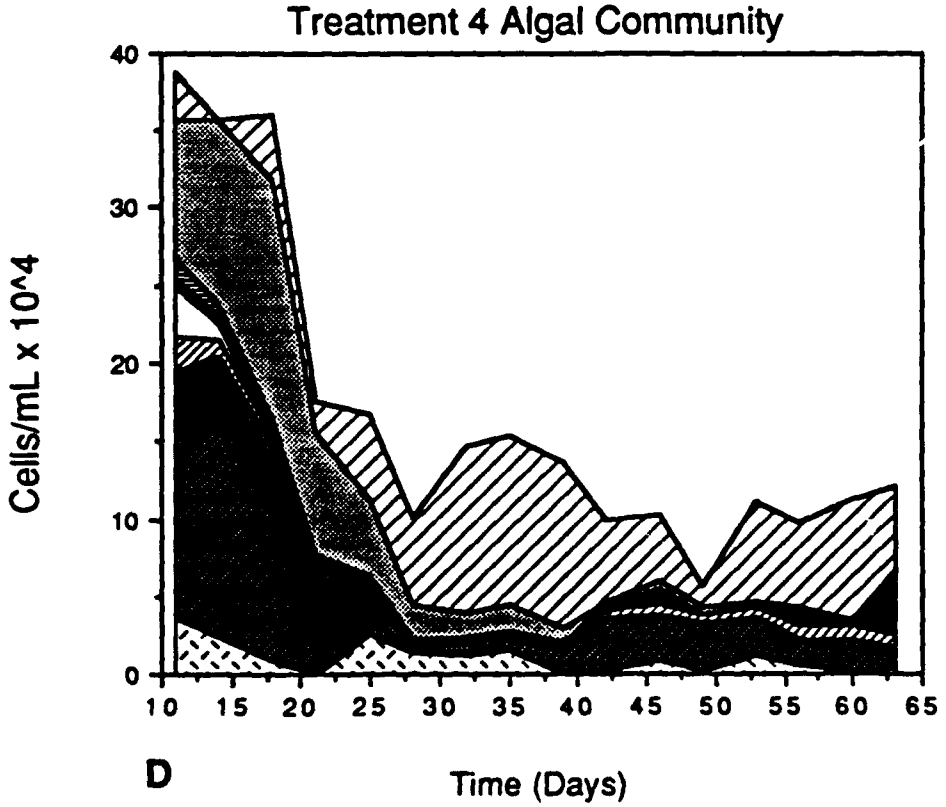



Treatment 1 (Non-Dosed) Algal Community
Cells/mL x 10^4
40
30
20
10
0
10 15 20 25 30 35 40 45 50 55 60 65
A
Time (Days)
Lyngbya
Scenedesmus
Selanastrum
Stigeoclonium
Ulothrix
Chlorella
Chlamydomonas
Ankistrodesmus
Anabena

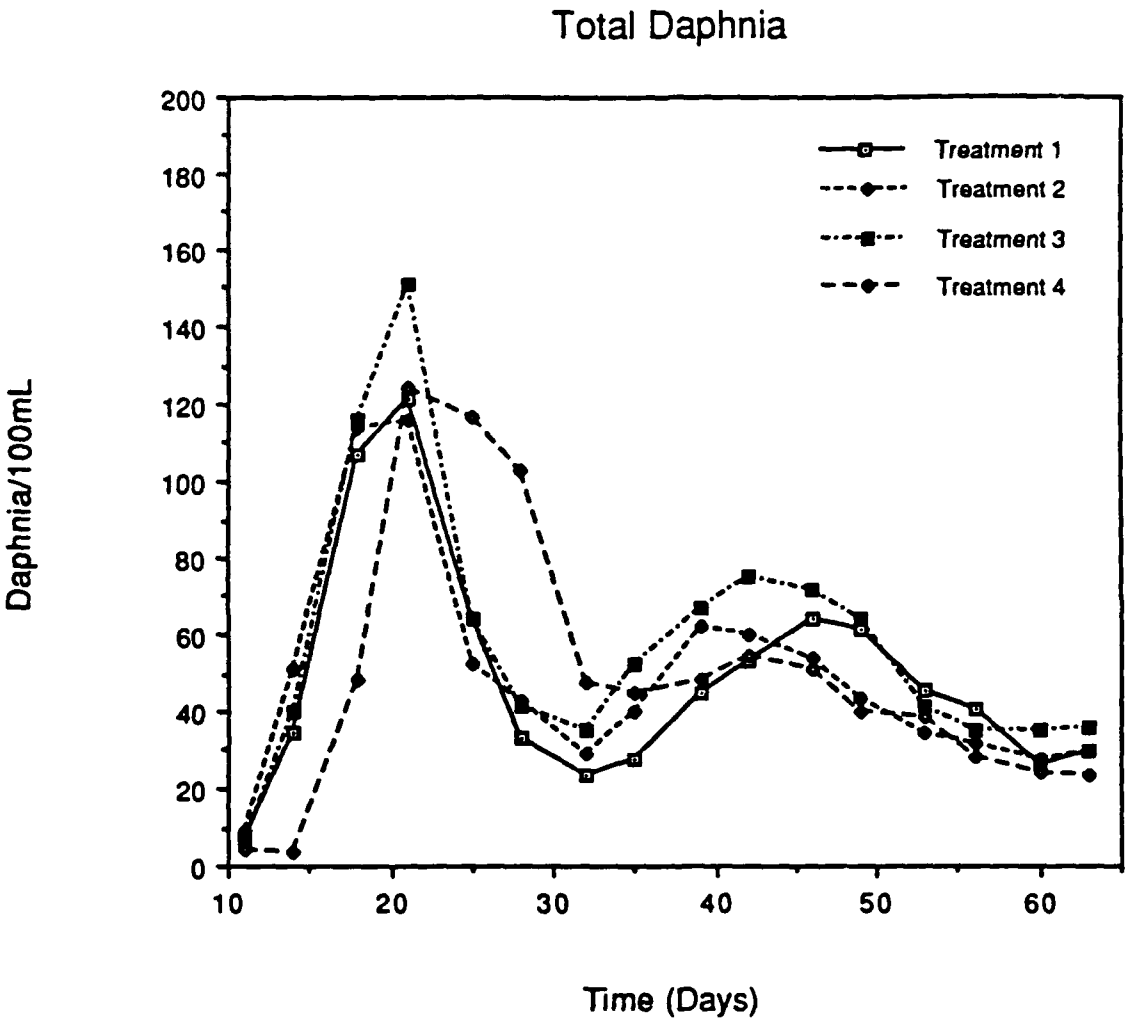

Treatment 2 Algal Community
Cells/mL x10^4
40
30
20
10
0
10 15 20 25 30 35 40 45 50 55 60 65
B
Time (Days)

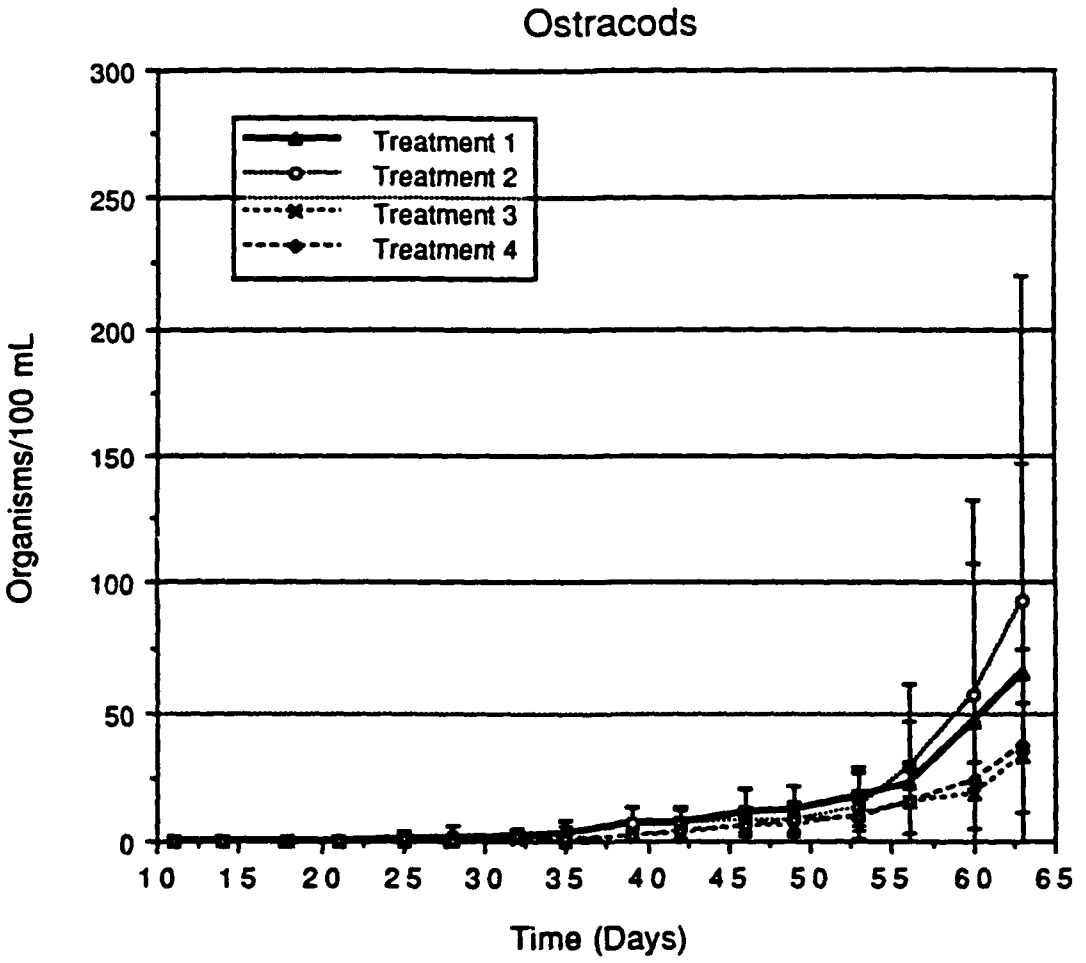

Treatment 3 Algal Community
Cells/mL x 10^4
40
30
20
10
0
10 15 20 25 30 35 40 45 50 55 60 65
Time (Days)

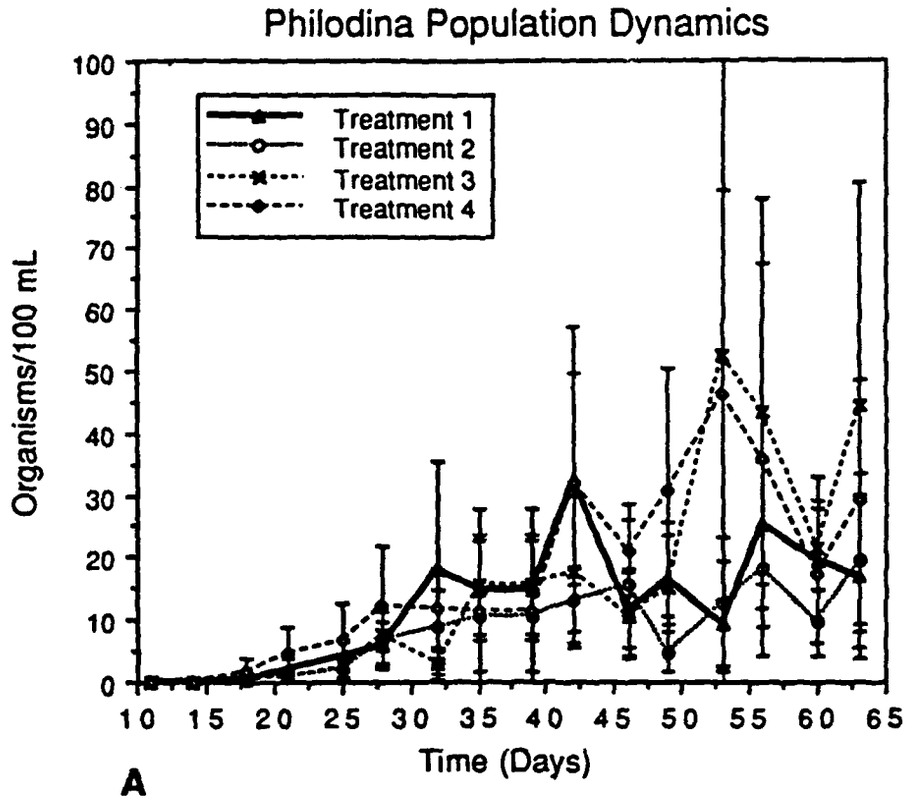
C

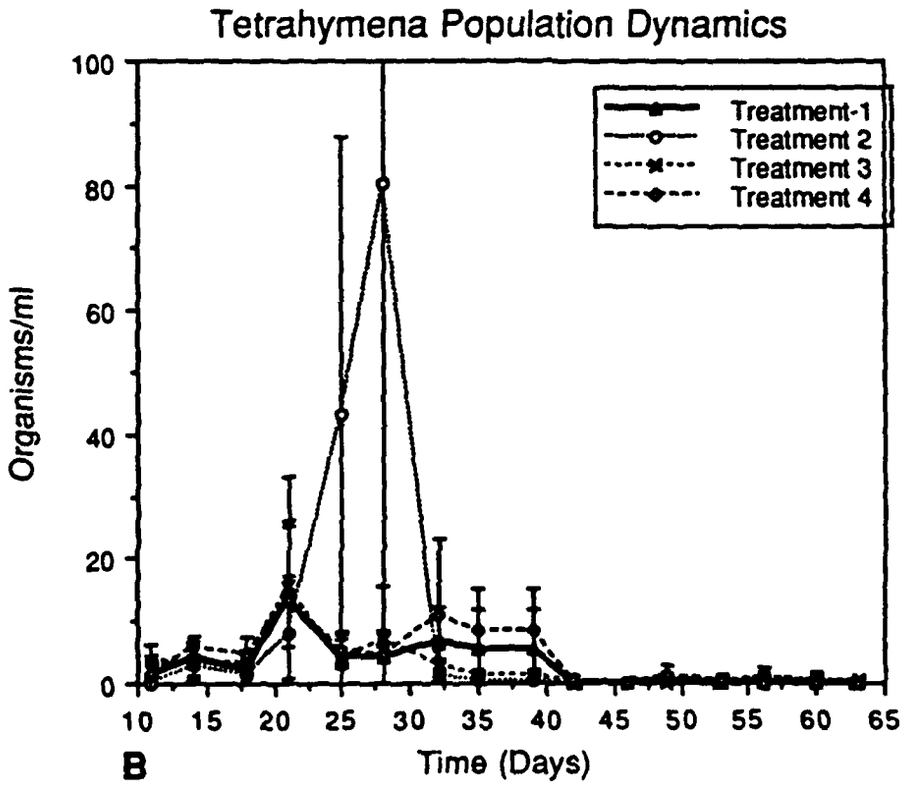

Treatment 4 Algal Community
Cells/mL x 10^4
40
30
20
10
0
10 15 20 25 30 35 40 45 50 55 60 65
Time (Days)

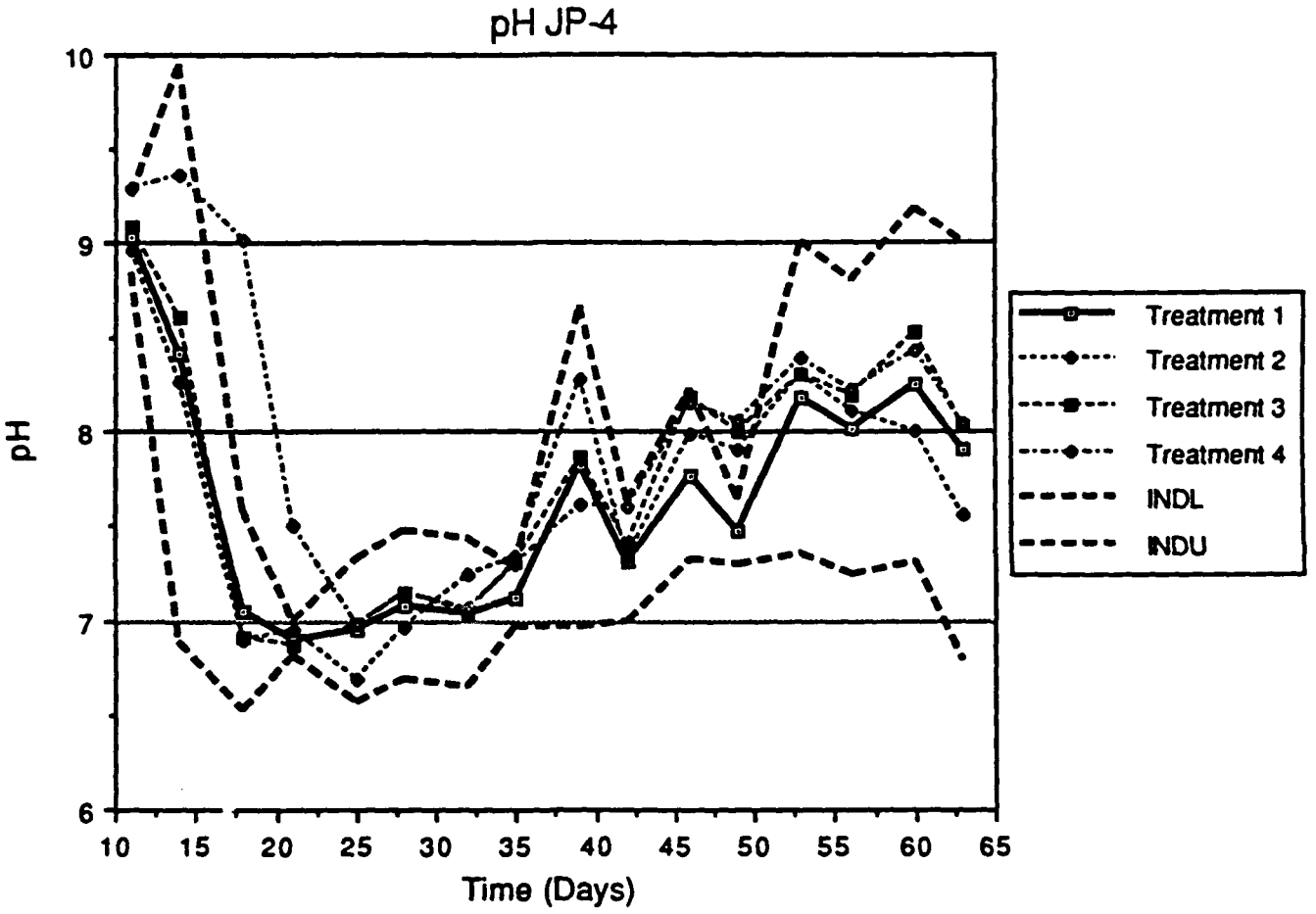
D

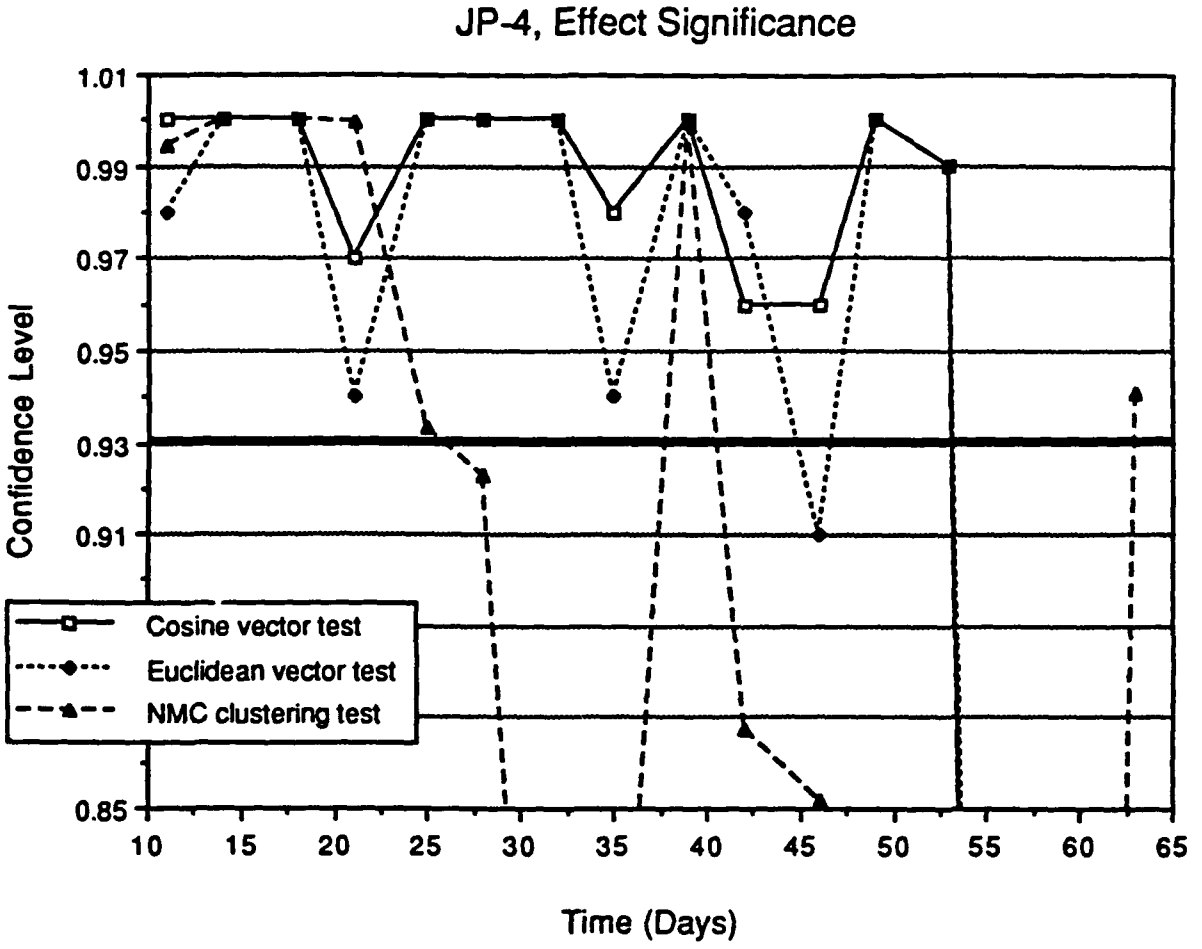

Total Daphnia
Daphnia/100mL
Time (Days)
Treatment 1
Treatment 2
Treatment 3
Treatment 4
0
20
40
60
80
100
120
140
160
180
200
10
20
30
40
50
60

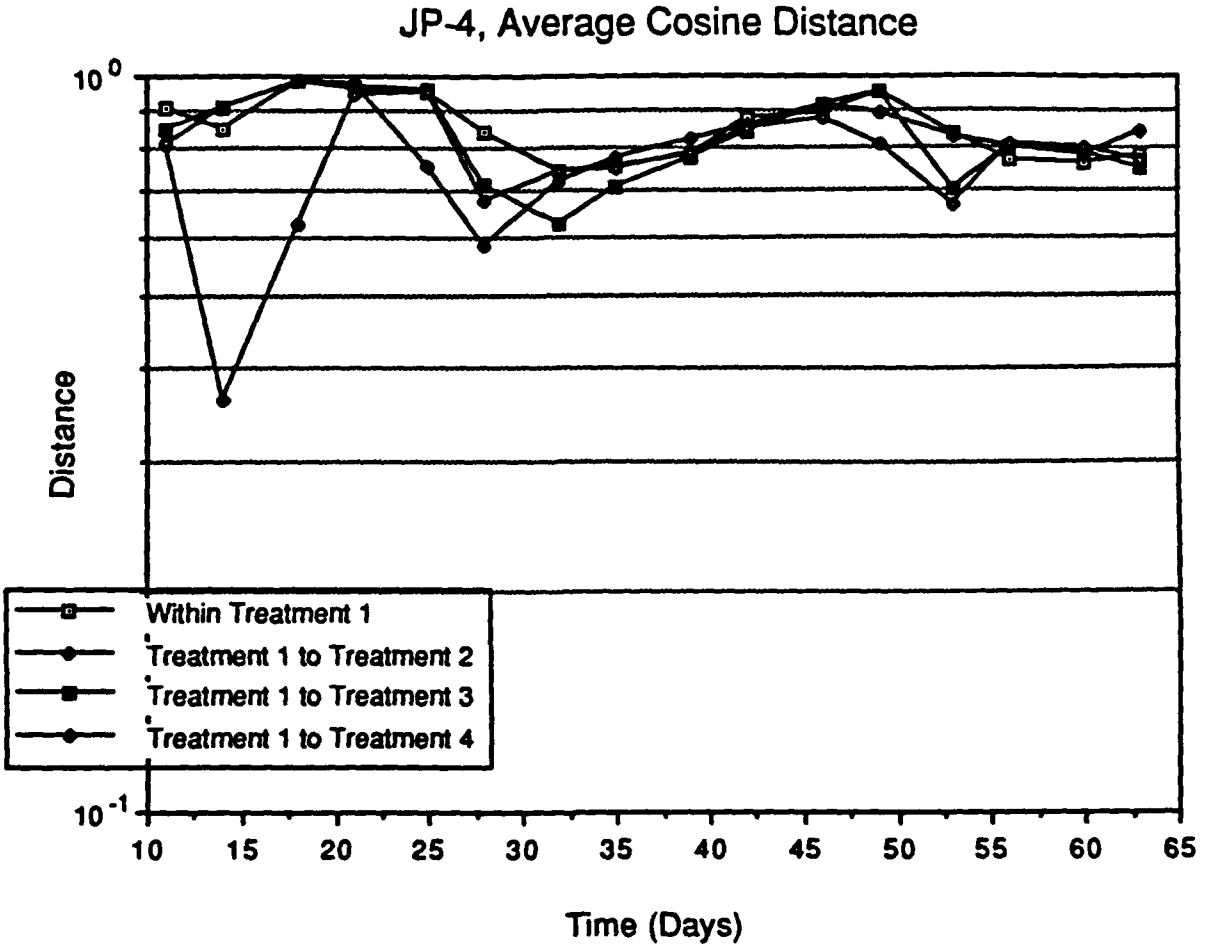

Ostracods
Treatment 1
Treatment 2
Treatment 3
Treatment 4
Organisms/100 mL
300
250
200
150
100
50
0
10
15
20
25
30
35
40
45
50
55
60
65
Time (Days)

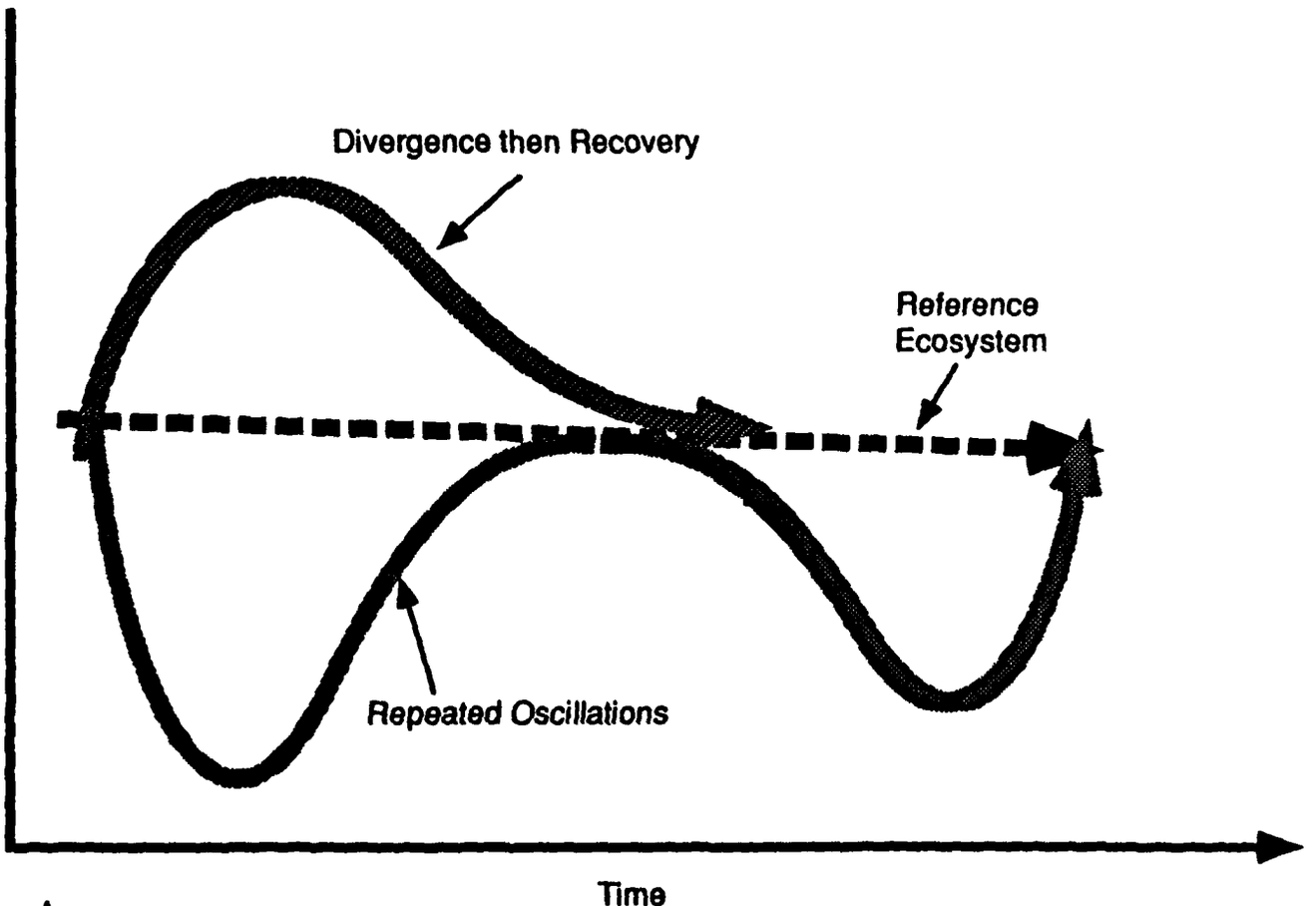

Philodina Population Dynamics
Treatment 1
Treatment 2
Treatment 3
Treatment 4
Organisms/100 mL
Time (Days)
A


Tetrahymena Population Dynamics
Treatment-1
Treatment 2
Treatment 3
Treatment 4
Organisms/ml
Time (Days)
B


pH JP-4
pH
10
9
8
7
6
10
15
20
25
30
35
40
45
50
55
60
65
Time (Days)
Treatment 1
Treatment 2
Treatment 3
Treatment 4
INDL
INDU

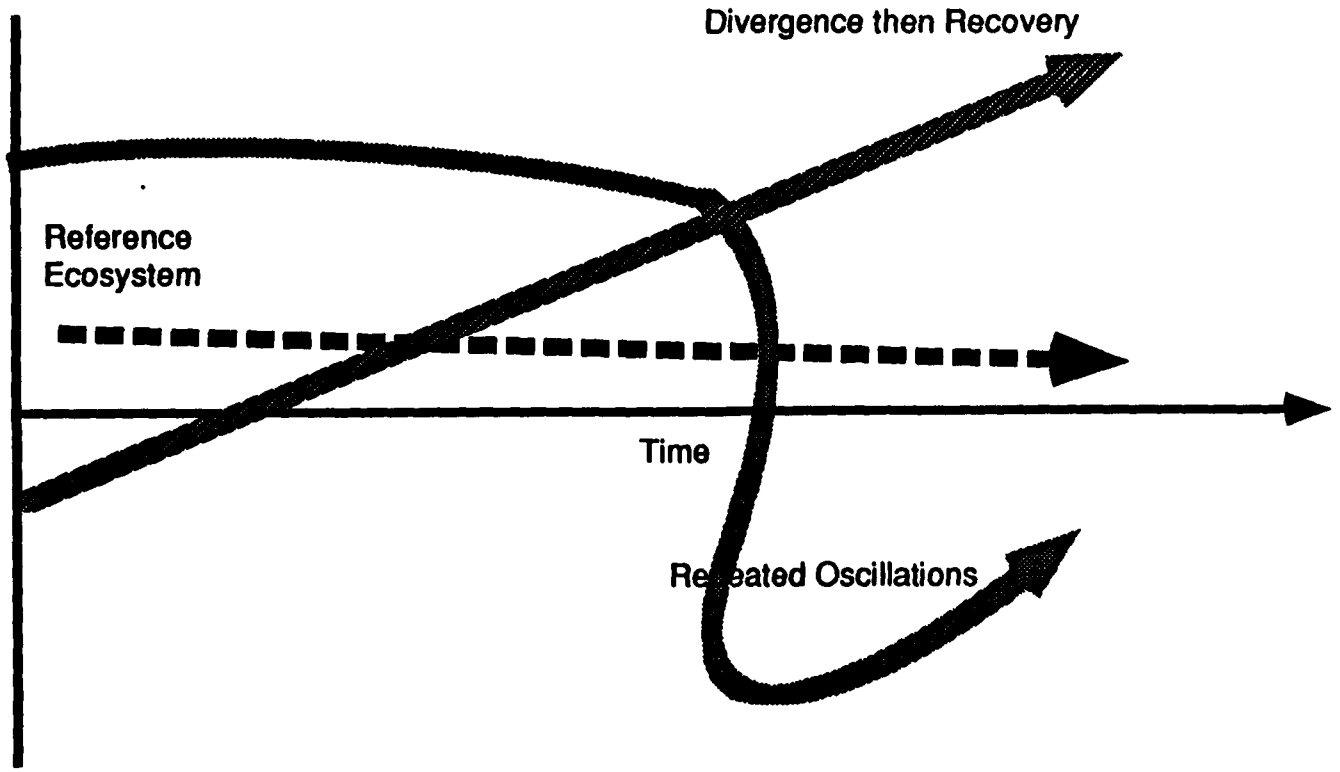

JP-4, Effect Significance
1.01
0.99
0.97
0.95
0.93
0.91
0.85
Confidence Level
10
15
20
25
30
35
40
45
50
55
60
65
Time (Days)
Cosine vector test
Euclidean vector test
NMC clustering test


JP-4, Average Cosine Distance
10^0
10^{-1}
Distance
Within Treatment 1
Treatment 1 to Treatment 2
Treatment 1 to Treatment 3
Treatment 1 to Treatment 4
10
15
20
25
30
35
40
45
50
55
60
65
Time (Days)


Divergence then Recovery
Reference
Ecosystem
Repeated Oscillations
Time

A

Top View


Divergence then Recovery
Reference
Ecosystem
Time
Repeated Oscillations

B

Application of Multivariate Techniques to Endpoint Determination, Selection and Evaluation in Ecological Risk Assessment

Wayne G. Landis[1], Geoffrey B. Matthews[2], Robin A. Matthews[1] and Anne Sergeant[3]

[1]Institute of Environmental Toxicology and Chemistry, Huxley College, Western Washington University, Bellingham, WA 98225

[2]Department of Computer Science, Western Washington University, Bellingham, WA 98225

[3]Anne Sergeant, ORD, U.S. EPA, 401 M Street, S.W. (RD-689), Washington DC 20460

Running Title: Multivariate Risk Assessment

Corresponding Author: Wayne G. Landis, IETC, Huxley College, Western Washington University, Bellingham, WA 98225-9079, 206-647-6136

Abstract: Ecological risk assessment has evolved so that the interaction among the components is now an implicit assumption. Unlike single species based risk assessments, it is often crucial in environmental or ecological risk assessments to be able to describe a system with many interacting components. In addition, some quantifiable description of how different biological communities are upon the addition of a toxicant or some other stressor is required to adequately describe risk at the ecosystem level. Three methods have been applied at the ecosystem level, the mean strain measurement used by K. Kersting, the state space analysis pioneered by A.R. Johnson, and the nonmetric clustering developed by G. Matthews for ecological datasets and for analysis of Standardized Aquatic Microcosm data. Each method has direct application to the description of an effected ecosystem without reliance upon a single and specific and perhaps misleading endpoint. Each also can assign distance or probability measures in order to compare the control to treatment groups. Nonmetric clustering (NMC) has the advantage of not attempting to combine different types of scales or metrics during the multivariate analysis and is robust against interference by random variables. Application of these methodologies into an ecological risk assessment should have the benefit of combining large interactive datasets into distinct measures to be used as a measure of risk and as a test of the prediction of risk. The primary impact of these methods may be in the selection and interpretation of assessment and measurement endpoints.

Much recent debate in toxicological studies has focused on appropriate endpoints for tests. Nonmetric clustering and other multivariate techniques should aid in the selection of these endpoints in ways meaningful at the ecosystem level. We suggest that the search for assessment and measurement endpoints be left to the appropriate multivariate computation algorithms in the case of multispecies situations. Application of these methods in the verification, validation process of risk assessment will prove to check the selection of endpoints during modeling exercises and to improve the presentation of assessment criteria.

Key Words: Risk assessment, multivariate statistics, nonmetric clustering, measurement and assessment endpoints, artificial intelligence.

Ecological Risk Assessment Defined

Ecological risk assessment is essentially the art of extrapolating from relatively straight-forward information on how toxic a compound is to specific organisms to how complex assemblages of organisms will respond to the toxin in their natural environment. The traditional approach to ecological risk assessment was developed by the National Academy of Science (NAS) using a human health effects paradigm. The NAS model is described in detail in *Risk Assessment in the Federal Government: Managing the Process* (1), also known as the "red book." The NAS approach uses a four-point approach:

a) The initial **hazard identification**, which determines whether a chemical is capable of causing adverse health effects. This conclusion is based on laboratory animal studies and, where available, human data;

b) The **dose-response assessment**, which characterizes the relationship between the chemical dose and the incidence of adverse health effects in the exposed population;

c) The **exposure assessment**, which measures or estimates the intensity, frequency, and duration of human exposure to a chemical, or estimates hypothetical exposure; and

d) The **risk characterization**, which combined the dose-response and exposure assessments. This final step evaluates the uncertainties in the previous analyses and provides an estimate of the likelihood of adverse effects under the stated conditions.

The NAS paradigm was developed to assess the risks of chemicals to human health, and while many of its principles can be implemented directly in ecological risk assessment, it falls short when applied to non-chemical stressors or interdependent organisms. Furthermore, it does not even begin to address the links between organisms and their environment. Hazard identifications are complicated by the many metabolic and degradation pathways available in the environment. Changes in these pathways can occur naturally, as a result of spatial and temporal changes in species assemblages, but can also be induced as a result of the introduction of a xenobiotic. Exposure assessments are complicated by the extraordinary array of species present at the exposure sites. The species composition also changes as a result of natural forces (seasonality, stochastic extinctions, migrations, etc.) or the introduction of a xenobiotic. Because of this, ecological risk assessment must be recognized as being fundamentally different from human health risk assessments (2).

Ecological Risk Assessment Models - Review of the USEPA Framework

Many of the difficulties in applying the traditional risk assessment paradigm to ecosystems have been addressed in the recent formulation of a *Framework for Ecological Risk Assessment* (3) (Figure 1). Among the novel features of this framework is the integration of exposure and hazard assessment to reflect the interactions that occur in ecological systems. Also innovative is the inclusion of a Data Acquisition, Verification and Monitoring process within the framework. The key however, is the selection

of assessment and measurement endpoints to make the assignment of risk representative of the system under protection.

The USEPA Framework includes three steps: problem formulation, analysis, and risk characterization.

Problem formulation is the process that evaluates the characteristics of the stress-inducing agent (e.g., toxin). It also identifies the ecosystem that may be at risk, and identifies possible ecological effects. This information is used to select the ecosystem components or attributes of concern (the assessment endpoints) and to determine the best ways to describe this component or attribute (measurement endpoints). Finally, the assessor prepares a conceptual model that describes the ways in which the stressor could interact with the ecosystem and the likely effects of such an interaction. Problem formulation is not specifically discussed in the NAS paradigm, but in current practice these issues are addressed during planning.

The **analysis** phase contains two components: **characterization of exposure** and **characterization of ecological effects**. The exposure characterization determines stressor distribution, characterizes receptors, and quantifies stressor release, migration, and fate. The effects characterization evaluates effects data and response data such as stressor-response analysis (akin to the dose-response assessment described above), the relationship between endpoints, and evidence of causality. This phase is analogous to the hazard identification, dose-response and exposure assessment components of the NAS paradigm.

The **risk characterization** component differs little from its counterpart in the NAS paradigm. It tests the hypotheses developed in the conceptual model described in Problem Formulation by synthesizing information about the stressor and receptor from various sources and describing the supporting evidence for (and uncertainty associated with) conclusions. It also provides some indication of the likelihood of effects occurring and describes the ecological significance of any predicted risk.

Endpoint Selection-Ecological Risk Assessment

Endpoints (assessment and measurement) are the keystones of an ecological risk assessment as every other parameter in the process is predicated upon these terms. An assessment endpoint must be something specific and quantifiable such as "maintenance of sport fish populations" or "desertification" or "eutrophication." Values such as "ecosystem health" have little meaning (2) and cannot be easily described. Sometimes it is not possible to examine the assessment endpoint directly--for example, one cannot collect bald eagle livers and analyze them for enzyme induction. In this case, measurement endpoints are used to describe the organism or entity of concern. Continuing with the bald eagle example, one may wish to examine contaminant concentrations in the eagles' food and compare them to laboratory dose-response data, observe their feeding habits and construct exposure scenarios, and review liver-enzyme data from other eagles (in captivity or found dead) or other birds of prey to arrive at

conclusions about enzyme induction in local eagles. In the ecosystem sense, measures of species number, abundance or energy flow would be analogous.

The USEPA Framework recommends that assessment endpoint selection consider 1) ecological relevance, 2) policy goals and societal values, and 3) susceptibility to the stressor. To ensure that ecological relevance is addressed, one must have some *a priori* knowledge of the ecosystem of interest and the relationships between its components. Science must not take a back seat to policy and societal values, but communication between the risk assessor and risk manager is critical to ensure scientific integrity and satisfy policy needs. Finally, the strongest assessment endpoints are both affected by the stressor and sensitive to a specific type of effect caused by that stressor.

Measurement endpoints should be selected on the basis of how well they represent assessment endpoints. Practicality and consistency with exposure scenarios often determine the initial range of possibilities. Measurement endpoints must be correlated with or useful for inferring changes in assessment endpoints (4). To the extent possible, they should be selected for appropriate diagnostic ability, signal-to-noise ratio, sensitivity, and response time. Ideally, measurement endpoints also provide information about indirect effects such as toxicity to an organism upon which the species of interest preys or nutrient cycle inhibition reducing survivorship of fingerlings.

An ecological risk assessment is only as good as the data upon which it is based. Thus, data acquisition is an integral part of the risk assessment process. Endpoints can and generally should change with time. At any stage in ecological risk assessment, new data may reveal that a particular endpoint should be added or removed, or that it no longer provides relevant information. For example, tree seedling success may be an important measure in managed ecosystems or when bare or disturbed soil is being colonized, but it provides little information about old-growth forests. Similarly, a measure of biomass in an aquatic system may provide a good indication of overall productivity, but it probably will not contain enough information to determine whether a balanced assemblage of functional groups (shredders, filter-feeders, etc.) exists. Preliminary data needs should be outlined during the Problem Formulation and refined as needed during the rest of the risk assessment process. For example, the assessor may discover that the assessment endpoint initially selected is affected less by the stressor being evaluated than by other causes, such as widespread habitat loss or overfishing--this may require selection of another assessment endpoint. Similarly, as the assessment progresses, it may become evident that additional measurement endpoints are needed. Increasingly, the use of multivariate data analysis is being called upon to assist in identifying appropriate endpoints for ecological risk assessments.

Importance of Multivariate Data in Ecological Risk Assessments

One important feature of ecological risk assessments is that they generally must rely on multivariate data to identify natural and toxicant-induced patterns. This is a result of the multidimensional nature of ecosystems; the Hutchinsonian idea of organisms and populations residing in a n-dimensional

hypervolume is the basis of current niche theory (5). The n-dimensional hypervolume is the ecosystem with all its components as perceived by the population. The variability of these parameters over time as well is used to account for the variety of species within the ecosystem system (6,7,8). Applications of resource competition models have been proposed for evaluating even single-species toxicant effects (9). Therefore, in order to begin to describe an ecosystem's response to perturbation, we must recognize the system's multidimensional nature.

Our essential goal in multivariate data analysis is to identify ecologically relevant patterns in the data set. This is true regardless of whether our ultimate goal is to develop an ecological risk assessment or to evaluate naturally occurring changes in the ecosystem. However, until recently, the data reduction tools available to aid our analyses have consisted primarily of simple graphs (lots of them), simple statistical tests done repeatedly to accommodate all of the measured parameters, and a few truly multivariate statistical tests that generated useful but esoteric results. For example, analysis of variance (ANOVA) is the classical method to examine single variable differences from control groups or reference sites. However, in multivariate data, there are problems with Type II errors. Furthermore, it is difficult to display and assimilate the many ANOVA results that are generated from a multivariate data set. Conquest and Taub (10) developed a method to overcome some of these problems by generating intervals of non-significant difference for a single variable measured repeatedly over time. This method corrects for the likelihood of a Type II error and produces a visual display of significant vs. nonsignificant differences that is easily graphed. The major drawback to this method is that it only portrays changes in single variables over time.

Multivariate methods have proved promising as a method of incorporating all of the dimensions of an ecosystem. One of the first to be used in toxicology was the calculation of ecosystem strain developed by Kersting (11,12,13,14) for relatively simple (three species) microcosms. At about the same time, Johnson (15,16) developed a multivariate clustering algorithm to map the n-dimensional coordinates of an ecosystem and used the distance between these systems as a measure of divergence from the control. Both of these methods have the advantage of examining the multispecies test systems as a whole and can track such process as succession, recovery and the deviation of a system due to an anthropogenic input. Their major disadvantage, which is also a disadvantage with most conventional multivariate statistical techniques, is that all of the data are incorporated without regard to the metric (unit of measurement) or relative value of a variable toward identifying patterns in the data set ("noisy" or random data are included along with the rest). It can be difficult to reconcile variables such as pH with a 0-14 metric to the numbers of bacterial cells per mL, where low numbers are in the 10^6 range. Along the same lines, data that vary randomly and have large metrics may overwhelm the statistical computations and mask the importance of highly correlated variables with small metrics.

Ideally, multivariate statistical tests used for evaluating complex data sets, whether the goal is to develop an ecological risk assessment or not, will have the following characteristics:

a) It will not combine counts from dissimilar taxa by means of sums of squares, or other ad hoc mathematical techniques, as in the Euclidean and cosine distance measures;

b) It will not require transformations of the data, such as normalizing the variance;

c) It will work without modification on incomplete data sets;

d) It will work without further assumptions on different data types (e.g., species counts or presence/absence data);

e) The Significance of a taxon to the analysis will not be dependent on the absolute size importance with common taxa, and taxa with a large, random variance will not automatically be selected to the exclusion of others;

f) It will provide an integral measure of "how good" the clustering is, i.e. whether the data set differs from a random collection of points; and

g) It will, if appropriate, identify a subset of the taxa that serve as reliable indicators of the physical environment.

Although we have now defined the ideal characteristics of a multivariate system, none is of course perfect. However, a method borrowed from the Artificial Intelligence (AI) tradition meets a large proportion of the above design criteria.

Nonmetric Clustering and Association Analysis

Unlike the more conventional multivariate statistics, nonmetric clustering is an outgrowth of artificial intelligence and a tradition of conceptual clustering. In this approach, an accurate description of the data is only part of the goal of the statistical analysis technique. Equally important is the intuitive clarity of the resulting statistics. For example, a linear discriminant function to distinguish between groups might be a complex function of dozens of variables, combined with delicately balanced factors. While the accuracy of the discriminant may be quite good, use of the discriminant for evaluation purposes is limited because humans cannot perceive hyperplanes in highly dimensional space. By contrast, conceptual clustering attempts to distinguish groups using as few variables as possible, and by making simple use of each one. Rather than combining variables in a linear function, for example, conjunctions of elementary "yes-no" questions could be combined: species A greater than 5, species B less than 2, and species C between 10 and 20. Numerous examples throughout the artificial intelligence literature have proven that this type of *conceptual* statistical analysis of the data provides much more useful insight into the patterns in the data, and is often more accurate and robust. Delicate linear discriminants, and other traditional techniques, chronically suffer from overfitting, particularly in highly dimensioned spaces. Conceptual statistical analysis attempts to fit the data, but not at the expense of a simple, intuitive result.

Applications of Nonmetric Clustering and Association Analysis

A detailed description of our multivariate methods, including nonmetric clustering and association analysis is in Appendix A. As examples of the usefulness of multivariate methods in general, and nonmetric clustering in particular, we will use examples of field evaluations and toxicity tests conducted over the last 3 years. Insights into the utility of these methods, the dynamics of even straightforward microcosm systems, and the importance of measurement variables have been the results of these studies.

Field Studies

Before we can determine whether a toxin has affected a group of organisms or the dynamics of an ecological community, we must first determine what types of changes would occur that are independent of the toxin. In field situations, this is usually attempted by using a reference site, monitoring the changes that occur at that site, and comparing this with the changes that occur in organisms at the "treatment" site.

However, one of the most difficult analytical challenges in ecology is to identify patterns of change in large ecological data sets. Often these data are not linear, they rarely conform to parametric assumptions, they have incommensurable units (e.g., length, concentration, frequency, etc.), and they are incomplete (due to both sample loss and sampling design whereby different parameters are collected at different frequencies). These difficulties exist regardless of whether there are toxins present; the only difference is that with the presence of a toxin, we must try to separate the response to the toxin from the other changes that occur at the site(s).

We have compared several types of multivariate techniques to evaluate two types of ecological data, a limnological data set that included spatial and temporal changes in water chemistry and phytoplankton populations, and a stream data set that included spatial (longitudinal) and temporal changes in benthic macroinvertebrate species assemblages (17,18) . Our objective was to see whether the multivariate tests could identify obvious patterns involving the influences of stratification in the lake and the effects of substrate and water quality changes on stream macroinvertebrates. We used principal components analysis, hierarchical clustering (k-means with squared Euclidean or cosine of vectors distance measures), correspondence analysis, and nonmetric clustering to look for patterns in the data.

In both studies, nonmetric clustering outperformed the metric tests, although both principal components analysis and correspondence analysis yielded some additional insight on large-scaled patterns that was not provided by the nonmetric clustering results. However, nonmetric clustering provided information without the use of inappropriate assumptions, data transformations, or other data set manipulations that usually accompany the use of multivariate metric statistics. The success of these studies and techniques lead to the detailed examiniation of community dynamics in a series of two multispecies toxicity tests.

Multispecies Toxicity Testing

The multivariate methods described above have recently been used to examine a series of multispecies toxicity tests. Described below are the data analyses from two recently published tests using methodology derived from the Standardized Aquatic Microcosm (SAM) (ASTM E1366-91). The 64-day SAM-protocol previously has been described (19,20,21,22,23). Briefly, the microcosms were prepared by the introduction of ten algal, four invertebrate, and one bacterial species into 3L of sterile defined medium.

In the first example (24), the riot control material 1,4-dibenz oxazepine (CR) was degraded using the patented organism *Alcaligenes denitrificans denitrificans* CR-1 (*A. denitrificans* CR-1). *A. denitrificans* CR-1 was obtained using a natural inoculum set in an environment containing the microcosm medium T82MV containing the toxicant CR. After demonstrating the organisms ability to degrade the toxicant CR, a microcosm experiment was set up to investigate the ability of the microorganisms to degrade CR in an environment resembling a typical freshwater environment. Toxicity tests of the riot control material demonstrated that although *A. denitrificans* CR-1 eliminated the toxicity of a CR solution towards algae, toxicity did remain to *Daphnia magna.*

The SAM experiment was set up with a control group without the toxicant or *A. denitrificans* CR-1, a second group with only CR, a third group with only *A. denitrificans* CR-1, and the fourth group containing both the toxicant CR and the bacterium *A. denitrificans* CR-1. Conventional analysis demonstrated that the major impact was the increase in algal populations since both CR and the degradative products of the toxicant both inhibited the growth of the major herbivore, *D. magna.* The control group and the microcosms inoculated initially with *A. denitrificans* CR-1 were not distinguishable using conventional analysis.

As a first test of the use of multivariate analysis in the interpretation of multispecies toxicity tests, the data set used to analyze the CR microcosm experiment were presented in a blind fashion for analysis. Neither the purpose of the experiment or the experimental set up was provided for the analysis. Nonmetric clustering was used to rank variables in terms of contribution and to set clusters. Surprisingly, the analysis resulted in only two clusters being recognized, Control and *A. denitrificans* CR-1 treatments, and the CR and CR plus *A. denitrificans* CR-1 treatments. Variables important in assigning clusters were D. magna, Ankistrodesmus, Scenedesmus and NO_2. Obviously, the inclusion of the principal algal species in these experiments and the daphnia was not a surprise, but NO_2 had not been demonstrated as a significant factor in previous analysis. However, the species *A. denitrificans denitrificans* is classified for its denitrification ability (25).

The second major application of nonmetric clustering to the analysis of SAM data has been the investigation of the impact of the water soluble fraction (WSF) of the fuel Jet-A (26). Four treatment groups, control, 1, 5 and 15 percent WSF were used.

All of the multivariate tests (cosine distance, vector distance and nonmetric clustering) agree that a significant difference between treatment groups was observed through day 25. From day 28 to day 39, the effect diminished until there were no significant effects observable. However, significant effects were again observable from day 46 through day 56, after which they again disappeared for days 60 and 63.

In Figure 2, the average cosine distances within the control group and between the control group and each of the three treatment groups are plotted on a log scale. The initial, strong effect, from day 11 to day 25, is easily seen as a large distance from the treatment 1 (control) and treatment 2, together, to both treatment groups 3 and 4, initially, but then treatment 3 moves closer to the control. The period of no significant difference, from day 35 to day 46, is also clear. During the second period of significant difference, from day 49 to 59, a perfect dose-response for all three treatments is seen, with higher doses becoming more distant from the control. This dose-response relationship is consistently maintained over a period of eleven days, for four sampling dates, days 49, 53, 56, and 59. In general, a dose-response relationship like this was not observed earlier, although the magnitude of the distance was considerably greater.

Also of interest are the variables that best described the clusters and the stability of the importance of the variables during the course of the experiment. Table 1 lists the variables determined to be important in determining the clusters by importance for each sampling day as determined by nonmetric clustering. In general, the number of variables that were important was larger during the start of the test and lower at the end. In addition, a great deal of variability in rankings is apparent during the course of the SAM. The number of sampling dates when a variable was deemed important in cluster formation is listed in Table 2. Ankistrodesmus was the most consistent of the variables, being ranked in 12 out of the 16 sampling dates. Medium daphnia was also ranked often. However, variables like Ostracod and Philodina did not become important until later in the experiment.

The repeated oscillation of the dosed replicates compared to the controls were accounted for in two basic ways:

a reflection of the functioning of the community best described by parameters not directly sampled by the SAM protocol; or,

a repeated fluctuation in community structure initiated by the initial stress and that is visible as an undampened movement in the systems.

Until more data can be obtained, the cause-effect of the second oscillation can not be determined. However, the use of multivariate analysis detected an unexpected result, one providing a new insight into the dynamics of even the relatively simple laboratory microcosm.

Synthesis

Several other researchers have attempted to employ multivariate methods to the description of ecosystems and the impacts of chemical stressors. Perhaps the best developed approaches have been those of K. Kersting and A.R. Johnson.

Multivariate Descriptions of Microcosm Systems

Normalized Ecosystem Strain (NES) was developed by Kersting (11,13) as a means of describing the impacts of several materials to the three compartment microecosystems containing an autotrophic, herbivore and decomposer subsystems. These variables in the unperturbed control systems are used to calculate the normal operating range (NOR) of the microecosystem. The NOR is the 95 per cent confidence ellipsoid of the unperturbed state of a system. The center of the NOR is defined as the reference point for the calculation of the NES. The NES is calculated as the quotient of the Euclidean distance from a state to the reference state divided by the distance from the reference state to the 95 percent confidence (also called tolerance) ellipsoid, along the vector that connects the reference state to the newly defined state. A value of 1 or less indicates that the new state is within the 95 percent confidence ellipsoid, values greater than 1 indicate that the system is outside this confidence region.

Originally limited to ellipsoids, the use of Mahalonobis distances allows the use of more variables as the confidence ellipsoid can be transformed to a confidence or tolerance hypersphere. These ideas were examined using the microecosytem test method developed by Kersting for the examination of multispecies systems. In tests using a relatively straightforward multicompartment microcosm the sensitivity and strengths of this methods were observed. The sensitivity of the NES increased sensitivity as the number of variables used to describe the system increased (13). Another interesting observation was the increasing distance from the normal space of the system after a perturbation as measured by NES as time increased. This increasing distance indicates that the perturbed system is drifting from its original state. Kersting hypothesized that the system may even shift to a different equilibrium state or domain and that the system would remain there even after the release of the stressor.

Apparently as an independent development, A.R. Johnson (15) proposed the idea of using a multivariate approach to the analysis of multispecies toxicity tests. This state space analysis is based upon the common representation of complex and dynamic systems as an n-dimensional vector. In other words, the system is described at a specific moment in time as a representation of the values of the measurement variables in an n-dimensional space. A vector can be assigned to describe the motion of the system through this n-dimensional space to represent successional changes, evolutionary events, or anthropogenic stressors. The direction and position information form the trajectory of the state space and this can be plotted over time.

In the n-dimensional hypervolume that describes the placement and trajectory of the ecosystem it is possible to compare the positions of systems at a specified time. This displacement can be measured by

literally computing the distance from the systems and this displacement vector can be regarded as the displacement of these systems in space. This displacement vectors can be easily calculated and compared. Using the data generated by Giddings (27) in a series of classic experiments comparing results of the impacts of synthetic oil on aquarium and small pond multispecies systems, Johnson was able to plot dose response curves using the mean separation of the replicate systems. These plots are very reminiscent of dose-response curves from typical acute and chronic toxicity tests.

As summarized by Johnson, the strengths of this methodology are the objectivity for quantifying the behavior of the stressed ecosystem and the power of this methodology to summarize large amounts of data. As with the work of Kersting, this methodology allows the investigator to examine the stability of the ecosystem and the eventual fate of the system relative to the control treatment.

Another important application proposed by Johnson (16) was the use of multivariate analysis to identify diagnostic variables that can be applied in the monitoring of ecosystems. Diagnostic variables, if reliable in differentiating anthropogenically stressed systems from control systems would be extremely valuable in monitoring for compliance and in determining clean up standards. The use of such variables is justified due to the fact that decisions often have to be made with incomplete datasets due to technical difficulties, cost, and a general lack of knowledge. Techniques proposed for the determination of these variables included linear regression, discriminant analysis and visual inspection of graphed data. Johnson conducted a cost-benefit analysis using an ecosystem model that demonstrated under the condition of that model, the benefits of diagnostic variables. In the Discussion, Johnson proposes simulation modeling to attempt to find generallzed diagnostic variables that best describe the state space and trajectory of an ecosystem.

The major difficulty with the methods detailed above is the reliance on conventional metric statistics. Vector distances in an n-dimensional space including such disparate variables as pH, cells counts and nutrient concentrations are difficult to compare from one experiment to another. Another consideration is the fact that many of the variables may be compilations of others. Algal biomass is often calculated by using multiplying cell counts by an appropriate constant for each species. Species diversity and many indices of ecosystem health are similarly composited variables. As discussed in the pervious sections, the use of metric methods with nonmetric clustering may prove a useful combination.

Search for Relevant Assessment and Measurement Endpoints

The attempt by Johnson to derive diagnostic variables is an interesting approach. However, our current research indicates that identity of the variables that contribute the most to separating control treatment from dosed treatment groups change from sampling period to sampling period. The variables change in the SAM experiments, no doubt, in response to the successional trajectory of the system as nutrients become depleted. As nutrients become limiting and the ability of the system to exhibit large

differences in community structure become less, the metric measures do not exhibit the same magnitudes of separation. Nonmetric clustering does not seem to be as sensitive to these changes.

However, the search for diagnostic measures to indicate the displacement of an ecosystem may not be fruitless. Although the relative importance of the variables in the SAM experiments may change, there are often variables that are more critical during the earlier stages of the development of the microcosm and those that are more crucial in the latter stages. The variable Ostracods is generally more important in the latter half of the experimental series than in the latter stages. The crucial aspect is that the clustering algorithm is able to select ecosystem attributes that are the best in differentiating stressed versus non-stressed systems. Although expert judgment may be able to predict in some cases variables that could be considered important to measure, the clustering approach is rapid, consistent, and not biased.

Instead of defining Assessment Endpoints, it may be more practical to define an Assessment Baseline or hypervolume using variables that have been demonstrated to be important in past descriptions of these types of ecosystems. Defining the 95 percent confidence region may be a more accurate way of characterizing the problem than by using artificial constructs or individual assessment measurement endpoint combinations. Assignment of these confidence regions may also improve the quality and accuracy of environmental risk assessment. Another logical outcome is that these regions must be defined by the measurement endpoints (variables). Measurement endpoints are the means by which a system can be accurately placed and its trajectory defined in an n-dimensional coordinate system. Such a means of describing systems has already been proposed by Kersting. The confidence region used to calculate NES is static, but an accounting of the passage of such a system through the coordinate system should provide a region from which deviation can be measured. Comparing dosed treatment groups to a control group is essentially the corresponding exercise but using a control series of replicates instead of an *a priori* prediction to measure deviation from the Assessment Baseline hypervolumes.

Measurement endpoints are therefore operationally defined, in the context of this paper using a multivariate approach, as the variables the set the axes for the description of the system within the n-dimensional space. Data such as dose-response curves may play a part if they describe a relevant axes when used in a biomonitoring role. Dose response data, however, are not measurement endpoints by themselves, but are important in setting relevant system parameters. It is preferable to select measurement endpoints that are the lowest common denominator of the system that is capable of being measured. For example, pH is certainly the most direct measurement of hydrogen ion concentration available. Diversity and other indices of species number and community structure, however, are composites of species abundance data.

The Myth of Ecosystem Health and Measurement Indices

The use of indices such as diversity and the Index of Biological Integrity have the effect of collapsing the dimensions of the hypervolume in a relatively arbitrary fashion. Indices, since they are composited variables, are not true endpoints. The collapse of the dimensions that are composited to tends to eliminate crucial information, such as the variability and distribution of the organisms within a particular system. The mere presence of absence and the frequency of these events can be analyzed using techniques such as nonmetric clustering and preserves the nature of the dataset. A useful function was certainly served by the application of these methods, but the new methods of data analysis and compilation should serve to replace these approaches and preserve the underlying structure and dynamic nature of ecological systems.

Part of the attraction of using indices may result in the pervasive nature of the metaphor, ecosystem health. In a recent critical evaluation, Suter (2) dismissed ecosystem health as a misrepresentation of ecological science. Ecosystems are not organisms with the patterns of homeostasis determined by a central genetic core. Since ecosystems are not organismal in nature, health is a property that can not describe the state of such a system. The urge to represent such a state as health has lead to the compilation of variables with different metrics, characteristics and casual relationships. Suter suggests a better alternative would be to evaluate the array of ecosystem processes of interest, a process that is now possible given multivariate methods.

Future Developments

Modeling of ecosystems may play an even more important role as the ability to generate the Assessment Baseline hypervolumes increases. However, the critical aspect is that these models not only predict the outcomes of the species under protection or the fishery that must be preserved but also the values of the measurements that can be made in a field or laboratory situation. These predictions should also predict sampling variability and chaotic and stochastic variation. The development of such models would be a critical development in the formulation of risk assessment methodologies.

Development of such models should be made with the understanding that the probability of divergence from the control state or the Assessment Baseline hypervolume given enough time will be 1.00. Assessment goals should be defined with reasonable time periods.

A major difficulty in the exploitation of these methods is that the vector distances, and to some extent even the cosine distances are not transferable or comparable unless the variables measured are essentially the same with the same metrics. Systems with different descriptive parameters will by definition occupy a different volume of n-dimensional space, making comparisons difficult. Determining the relevant parameters to use a measurement endpoints *a priori* may be difficult if not impossible.

There are benefits that should evolve directly from the use of multivariate techniques. First, it should force the description of measurement and assessment endpoints in terms of acceptable variance in a

dynamic fashion with expected distributions or functionality. Probabilistic criteria will certainly evolve from these aspects.

As these criteria are developed, the recognition that ecosystems are unique in their basic nature and not amenable to descriptions that incorporate only one dimensionally with that dimension an arbitrary axis.

Finally, the use of multivariate techniques should enable the researcher and assessor the capability of using all of the data in the description of an ecosystem with the results presentable to a decision maker or risk manager. After all, it has proven feasible to portray the results of these analysis in terms of distance and probabilities.

Acknowledgments. This research is supported by United States Air Force Office of Scientific Research Grant No. AFOSR-91-0291 DEF. We would also like to thank G. Suter, S. Norton, J. Dulka and B. Weber for their thoughtful discussions and encouragement.

Appendix A. Multivariate Techniques

In the research described below, three multivariate significance tests were used. Two of them were based on the ratio of multivariate metric distances within treatment groups vs. between treatment groups. One of these is calculated using Euclidean distance and the other with cosine of vectors distance (28,29) (Figure 3). The third test used nonmetric clustering and association analysis (30). In the microcosm tests there were four treatment groups with six replicates, giving a total of 24. This example is used to illustrate the applications in the derivations that follow.

Treating a sample on a given day as a vector of values, $\vec{x} = \langle x_1, \ldots x_{17} \rangle$, with one value for each of the measured biotic parameters, allows multivariate distance functions to be computed. Euclidean distance between two sample points $\vec{x}$ and $\vec{y}$ is computed as

$$\sqrt{\sum_i (x_i - y_i)^2}$$

The cosine of the vector distance between the points $\vec{x}$ and $\vec{y}$ is computed as

$$1 - \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum y_i^2}}$$

Subtracting the cosine from one yields a distance measure, rather than a similarity measure, with the measure increasing as the points get farther from each other.

The within-between ratio test used a complete matrix of point-to-point distance (either Euclidean or cosine) values. For each sampling date, one sample point $\vec{x}$ was obtained from each of six replicates in the four treatment groups, giving a 24 x 24 matrix of distances. After the distances were computed, the ratio of the average within group metric (*W*) to the average between group metric (*B*) was computed (*W/B*). If the points in a given treatment group are closer to each other, on average, than they are to points in a different treatment group, then this ratio will be small. The significance of the ratio is estimated with an approximate randomization test (31). This test is based on the fact that, under the null hypothesis, assignment of points to treatment groups is random, the treatment having no effect. The test, accordingly, randomly assigns each of the replicate points to groups, and recomputes the *W/B* ratio, a large number of times (500 in our tests). If the null hypothesis is false, this randomly derived ratio will (probably) be larger than the *W/B* ratio obtained from the actual treatment groups. By taking a large number of random reassignments, a valid estimate of the probability under the null hypothesis is obtained as $(n+1)/(500+1)$, where n is the number of times a ratio less than or equal to the actual ratio was obtained (31).

In the clustering association test, the data are first clustered independently of the treatment group, using nonmetric clustering and the computer program RIFFLE (32). Because the RIFFLE analysis is naive

to treatment group, the clusters may, or may not correspond to treatment effects. To evaluate whether the clusters were related to treatment groups, whenever the clustering procedure produced four clusters for the sample points, the association between clusters and treatment groups was measured in a 4 x 4 contingency table, each point in treatment group i and cluster j being counted as a point in frequency cell ij. Significance of the association in the table was then measured with Pearson's X^2 test, defined as

$$\chi^2 = \sum_{ij} \frac{\left(N_{ij} - n_{ij}\right)^2}{n_{ij}}$$

where N_{ij} is the actual cell count and n_{ij} is the expected cell frequency, obtained from the row and column marginal totals N_{+j} and N_{i+} as

$$n_{ij} = \frac{N_{+j} N_{i+}}{N}$$

where $N = 24$ is the total cell count (33), and a standard procedure for computing the significance (probability) of χ^2 taken from (34).

REFERENCES

1. **National Research Council**. 1983. *Risk Assessment in the Federal Government.* National Academy Press, Washington, D.C.

2. **Suter, G.** 1993. A critique of ecosystem health: Concepts and indices. *Environ Tox. Chem.* in press.

3. **U. S. Environmental Protection Agency**. 1992. Framework for Ecological Risk Assessment. EPA/630/R-92/001.

4. **Suter II, G.W.** 1990. Endpoints for regional risk assessments. *Environmental Management.* 14:9-23.

5. **Hutchinson, G.E.** 1959. Concluding remarks. *Cold Spring Harbor Symposium on Quanitative Biology.* **22**:415-427.

6. **Hutchinson, G.E.** 1961. The paradox of the plankton. *Amer. Nat.* **95**:137-143.

7. **Richerson, P., R. Armstrong** and **C.R. Goldman**. 1970. Contemporaneous disequilibrium, a new hypothesis to explain the "paradox of the plankton". *Proc. Natl. Acad. Sci. U.S.A.* **67**:1710-1714.

8. **Tilman, D.** 1982. *Resource Competition and Community Structure.* Princeton University Press, Princeton, NJ, pp. 11-138.

9. **Landis, W.G.** 1986. Resource competition modeling of the impacts of xenobiotics on biological communities. In T.M. Poston and R. Purdy, eds., *Aquatic Toxicology and Environmental Fate: Ninth Volume, ASTM 921.* American Society for Testing and Materials, Philadelphia, PA, pp. 55-72.

10. **Conquest, L.L.** and **F.B. Taub.** 1989. Repeatability and reproducibility of the Standard Aquatic Microcosm: Statistical properties. In U.M. Cowgill and L.R. Williams, eds., *Aquatic Toxicology and Hazard Assessment: 12th Volume, ASTM STP 1027.* American Society for Testing and Materials, Philadelphia, PA, pp. 159-177.

11. **Kersting, K.** 1984. Development and use of an aquatic micro-ecosystem as a test system for toxic substances. Properties of an aquatic micro-ecosystem IV. *Int. Revue ges. Hydrobiol.* **69**:567-607.

12. **Kersting, K.** 1985. Properties of an aquatic micro-ecosystem V. Ten years of observations of the prototype. *Verh. Internat. Verein. Limnol.* **22**:3040-3045.

13. **Kersting, K.** 1988. Normalized ecosystem strain in micro-ecosystems using different sets of state variables. *Verh. Internat. Verein. Limnol.* **23**:1641-1646.

14. **Kersting, K.** and **R. van Wijngaarden.** 1992. Effects of Chloropyrifos on a microecosystem. **11**:365-372.

15. **Johnson, A.R.** 1988a. Evaluating ecosystem response to toxicant stress: a state space approach. *Aquatic Toxicology and Hazard Assessment: 10th Volume, ASTM STP* 971, W.J. Adams, G.A. Chapman and W.G. Landis Eds., American Society for Testing and Materials, Philadelphia, pp. 275-285.

16. **Johnson, A.R.** 1988b. Diagnostic variables as predictors of ecological risk. *Environmental Management* **12**:515-523.

17. **Matthews, R.A., G.B. Matthews** and **W.J. Ehinger.** 1991. Classification and ordination of limnological data: a comparison of analytical tools. *Ecological Modeling.* **53**:167-187.

18. **Matthews, G.B., R.A. Matthews** and **B. Hachmoller.** 1991. Mathematical analysis of temporal and spatial trends in the benthic macroinvertebrate communities of a small stream. *Canadian Journal of Fisheries and Aquatic Sciences.* **48**:2184-2190.

19. **Taub, F.B., A.C. Kindig** and **L.L. Conquest.** 1987. Interlaboratory testing of a standardized aquatic microcosm. In W.J. Adams, G.A. Chapman and W.G. Landis, eds., *Aquatic Toxicology and Hazard Assessment: 10th Volume, ASTM STP 971.* American Society for Testing and Materials, Philadelphia, PA, pp. 385-405.

20. **Taub, F.B., A.C. Kindig, L.L. Conquest** and **J.P. Meador.** 1988. Results of the interlaboratory testing of the Standardized Aquatic Microcosm protocol. In G. Suter and M. Lewis, eds., *Aquatic Toxicology and Hazard Assessment: Eleventh Symposium, ASTM.* American Society for Testing and Materials, Philadelphia, PA.

21. **Taub, F.B.** 1988. Standardized aquatic microcosm - development and testing. *Aquatic Ecotoxicology* II.

22. **Taub, F.B.** 1989. Standardized aquatic microcosms. *Environm. Sci. Technol.* **23**:1064-1066.

23. **Kindig, A.C. , L.C. Loveday** and **F.B. Taub.** 1983. Differential sensitivity of new versus mature synthetic microcosms to streptomycin sulfate treatment. In W.E. Bishop, R.D. Cardwell and B.B. Heidolph, eds., *Aquatic Toxicology and Hazard Assessment: Sixth Symposium, ASTM 802.* American Society for Testing and Materials, Philadelphia, PA, pp. 192-203.

24. **Landis, W.G., M.V. Haley** and **N.A. Chester.** 1993. The use of the standardized aquatic microcosm in the evaluation of degradative bacteria in reducing impacts to aquatic ecosystems. In W.G. Landis, J. Hughes and M. Lewis., eds., *In press, Environmental Toxicology and Risk Assessment: First Volume, ASTM STP -1179.* American Society for Testing and Materials, Philadelphia, PA.

25. **Matthews, G.** and **R. Matthews.** 1991. A model for describing community changes. *Proceedings,* Conference on Pesticides in Natural Systems, EPA 910/9-91-001, United States Environmental Protection Agency Region 10, Corvallis, OR.

26. **Landis, W.G., R.A. Matthews, A.J. Markiewicz, N.J. Hough** and **G.B. Matthews.** 1993. Multivariate analysis of the impacts of turbine fuel using a standard aquatic microcosm toxicity test. *J. Env. Sci.* In Press.

27. **Giddings, J.M., B.R. Parkhurst, C.W. Hehrs** and **R.E. Millemann.** 1980. Toxicity of a coal liquefaction product to aquatic organisms. *Bull. Environ. Contam. Toxicol.* **25**:1-6.

28. **Good, I.J.** 1982. An index of separateness of clusters and a permutation test for its significance. *J. Statist. Comp. Simul.* **15**:81-84.

29. **Smith, E.P. K.W. Pontasch** and **J. Cairns, Jr.** 1990. Community similarity and the analysis of multispecies environmental data: a unified statistical approach. *Water Res.* **24**:507-514.

30. **Matthews, G.B.** and **R.A. Matthews.** 1990. A model for describing community change. In *Pesticides in Natural Systems: How Can Their Effects Be Monitored? Proceeding of the Conference,* Environmental Research Laboratory/ORD, Corvallis, OR, EPA 9109/9-91/011.

31. **Noreen, E.W.** 1989. *Computer Intensive Methods for Testing Hypotheses.* Wiley-Interscience, New York, NY.

32. **Matthews, G.B.** and **J. Hearne.** 1991. Clustering without a metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **13**:175-184.

33. **Fienberg, S.E.** 1985. *The Analysis of Cross-Classified Categorical Data.* MIT Press, Cambridge, MA.

34. **Press, W.H., B.P. Flannery, A.A. Teukolsky** and **W.T. Vetterline.** 1990. *Numerical Recipes in C, the Art of Scientific Computing.* Cambridge University Press, New York, NY.

Tables

Table 1. Important Variables Ranked By Nonmetric Clustering For Each Sampling Date For The Jet-A SAM Toxicity Test. Some variables such as *Ankistrodesmus* were consistently important in determining group clusters throughout the experiment. Some of the variables such as **Ostracod** and Philodina were more important in the latter stages of the experiment. The order of importance of the variables often changed from day to day, with no one variable being common to each sampling date. The variables used as part of the overall analysis were: Anabaena, Ankistrodesmus, Chlamydomonas, Chlorella, Daphnia (Ephipia, Small Daphnia, Medium Daphnia, Large Daphnia), Hypotricha, Lyngbya, Miscellaneous sp., Ostracod (Cyprinotus), Philodina (Rotifer), Scenedesmus, Selenastrum, Stigeoclonium, and Ulothrix.

Day	Important Variables in Determining Clusters in Rank Order
11	M. Daphnia, Chlorella, Chlamydamonas, Ulothrix, S. Daphnia,Selanastrum,Scenedesmus
14	S. Daphnia, M. Daphnia-Selenastrum[1], Chlamydamonas, Chlorella, L. Daphnia, *Ankistrodesmus*
18	*Ankistrodesmus*, S. Daphnia, Chlorella, Chlamydamonas, Selanstrum, L. Daphnia
21	*Ankistrodesmus*, S. Daphnia, L. Daphnia-M. Daphnia, Scenedesmus
25	Scenedesmus, S. Daphnia, L. Daphnia, Chlorella, Philodina-M. Daphnia
28	*Ankistrodesmus*, L. Daphnia, Scenedesmus
32	S. Daphnia, M. Daphnia, *Ankistrodesmus*, Chlorella
35	*Ankistrodesmus*
39	M. Daphnia-Selenastrum, **Ostracod**-*Ankistrodesmus*
42	M. Daphnia, **Ostracod**, Scenedesmus
46	Scenedesmus, *Ankistrodesmus*, S. Daphnia. M. Daphnia
49	Chlorella, Philodina, *Ankistrodesmus*, Lyngbya
53	*Ankistrodesmus*, **Ostracod,** Chlorella
56	M. Daphnia-Scenedesmus, *Ankistrodesmus*, Lyngbya
60	Lyngbya, M. Daphnia, Philodina, Chlorella
63	Chlorella, *Ankistrodesmus*, Philodina, **Ostracod**

[1] Hyphen between variables denotes equal rank

Table 2. Variable According to Success in Determining Clusters as Defined by Nonmetric Clustering in the Jet-A SAM Experiments. Variables such as Ankistrodesmus and the Daphnia classes were important in the course of this study. Reliance on even these two variables would have been misleading in the determination of the second oscillation.

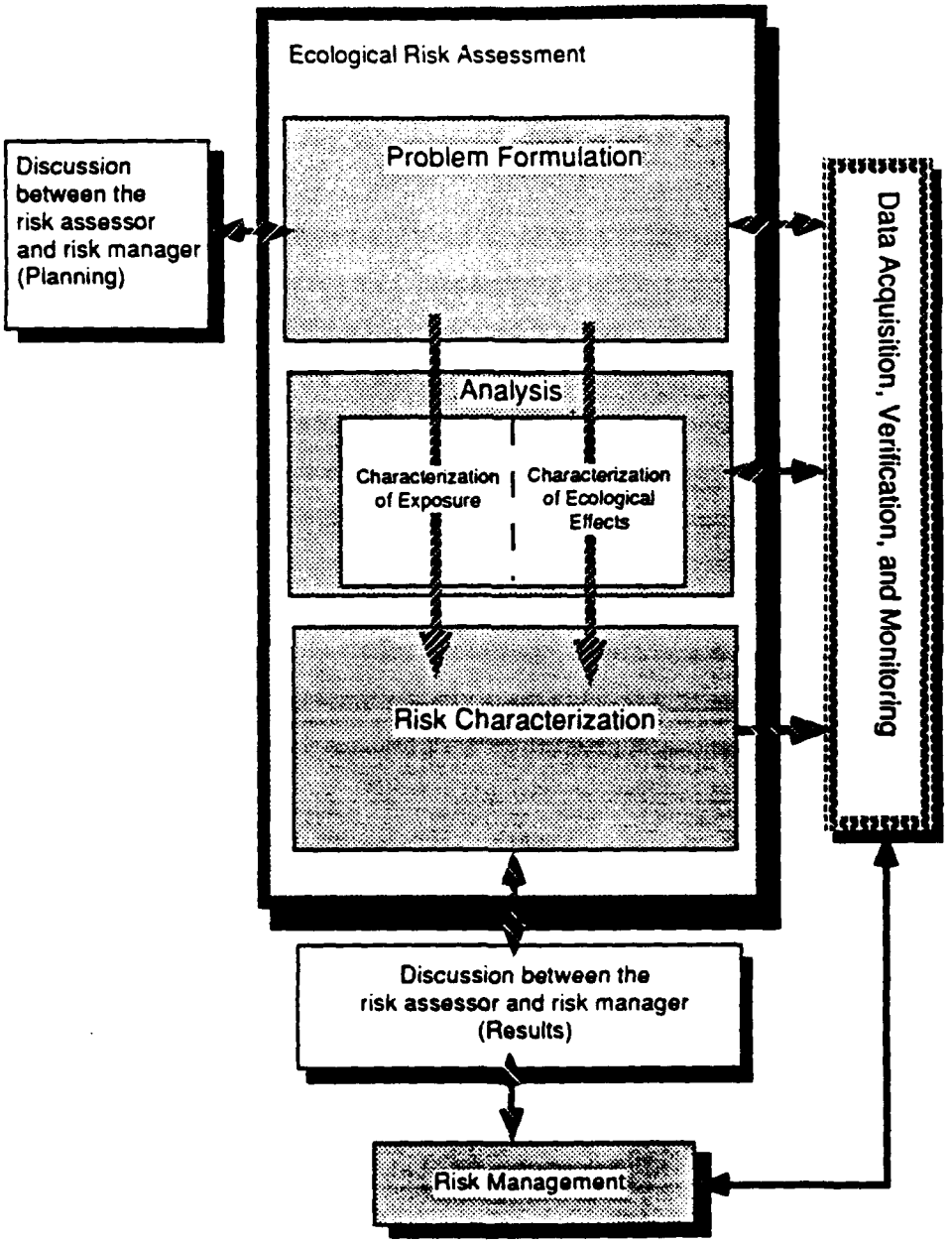
Variable	Ranked
Ankistrodesmus	12
M. Daphnia	11
Chlorella	9
Scenedesmus	7
S. Daphnia	6
L. Daphnia	5
Ostracod	4
Philodina	4
Selenastrum	4
Lyngbya	3
Ulothrix	1

Figures

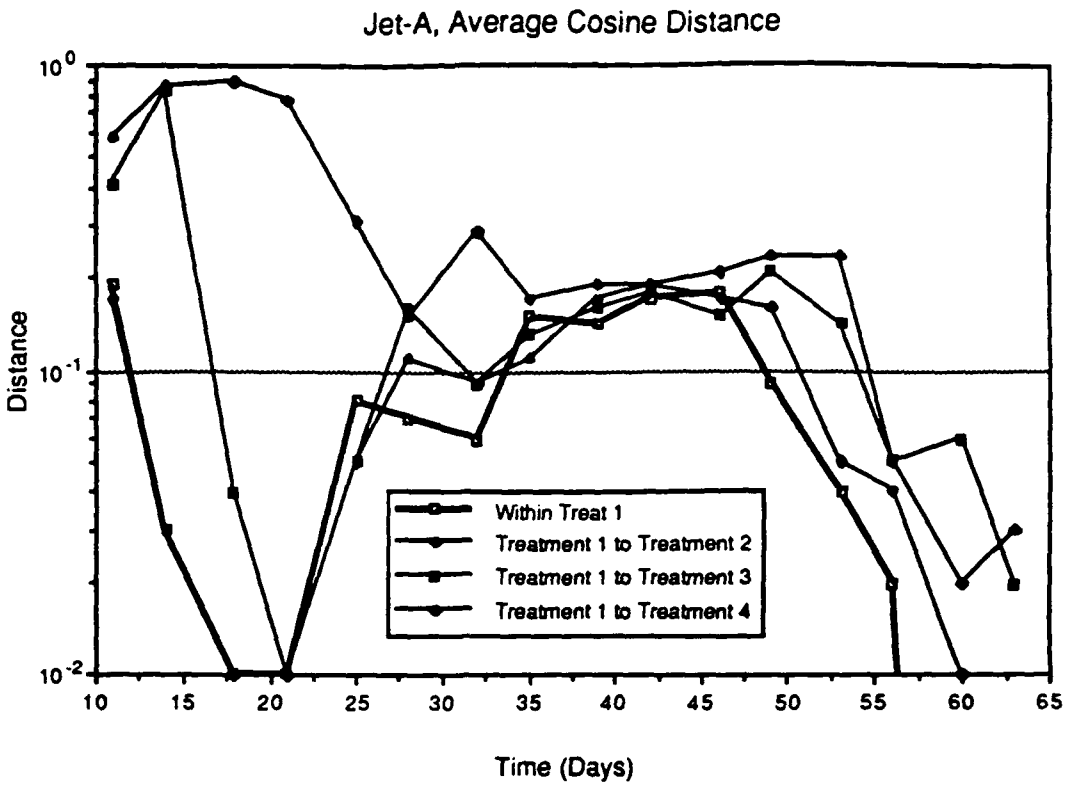
Figure 1. Schematic of the Framework for Ecological Risk Assessment (3). Especially important in the interaction between exposure and hazard and the inclusion of a data acquisition, verification and monitoring component. Multivariate analyses will have a major impact upon the selection or assessment and measurement endpoints as well as playing a major role in the data acquisition, verification and monitoring phase.

Figure 2. Multivariate analysis of the impact of Jet-A in the SAM test system. Figure 2A shows the Cosine distance from the control group to each of the treatments for each sampling day. Note that large differences are apparent early in the SAM. During the middle part of the 63 day experiment the distances between the replicates of Treatment 1, the control group, is as large as the distances to the treatment groups. However, later in the experiment the distances from the dosed microcosms to the control again increase. Significance levels of the three multivariate statistical tests for each sampling day are presented in Figure 2B. Note that there are two periods, early and late ones, where the clustering into treatment groups is significant at the 95 percent confidence level or above.

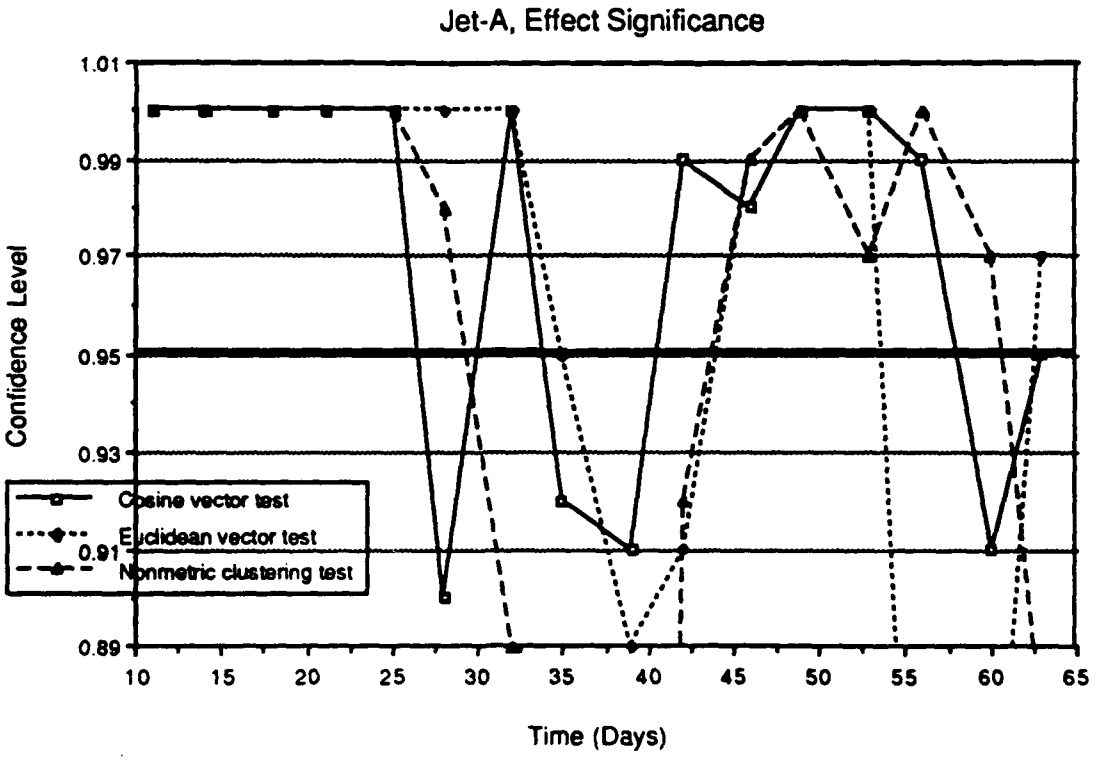
Figure 3. Measures of distance between clusters. Two of the commonly used measures of separation of clusters in a n-dimensional space are the cosine of the angle and the vector distance. Each method has advantages and disadvantages. In order to visualize the data as accurately as possible several measures should be employed.

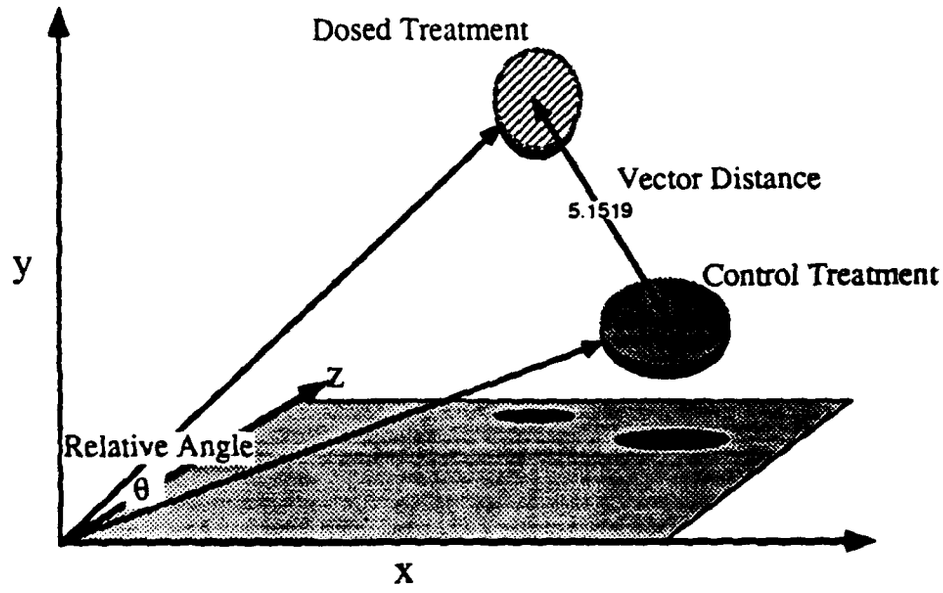

Ecological Risk Assessment
Problem Formulation
Analysis
Characterization of Exposure
Characterization of Ecological Effects
Risk Characterization
Discussion between the risk assessor and risk manager (Planning)
Data Acquisition, Verification, and Monitoring
Discussion between the risk assessor and risk manager (Results)
Risk Management

A


Jet-A, Average Cosine Distance
Distance
10^0
10^{-1}
10^{-2}
10
15
20
25
30
35
40
45
50
55
60
65
Time (Days)
Within Treat 1
Treatment 1 to Treatment 2
Treatment 1 to Treatment 3
Treatment 1 to Treatment 4

B


Jet-A, Effect Significance
Confidence Level
1.01
0.99
0.97
0.95
0.93
0.91
0.89
10
15
20
25
30
35
40
45
50
55
60
65
Time (Days)
Cosine vector test
Euclidean vector test
Nonmetric clustering test


Dosed Treatment
Vector Distance
5.1519
Control Treatment
y
z
Relative Angle
θ
x

Running head: Uncertainty Propagation in Risk Assessment

Address correspondence to:

Geoffrey Matthews

Computer Science Department, MS 9062

Western Washington University

Bellingham, Washington 98225

Office: (206) 676-3797

Fax: (206) 676-3044

Uncertainty Propagation in Risk Assessment

Geoffrey B. Matthews

Computer Science Department, MS 9062

Western Washington University

Bellingham, Washington 98225

Robin A. Matthews

Huxley College of Environmental Studies, MS 9079

Western Washington University

Bellingham, Washington 98225

Wayne G. Landis

Huxley College of Environmental Studies, MS 9079

Western Washington University

Bellingham, Washington 98225

James W. Hearne

Computer Science Department, MS 9062

Western Washington University

Bellingham, Washington 98225

Footnotes

(none)

Abstract

Risk assessment typically proceeds by successively combining various uncertain inferences into an overall probability. For example, in computing the potential effect on a target species, an extrapolation may have to be made from an acute test on a similar species. A test on white mice, for example, may be pressed into service to estimate effects on deer mice. The expected exposure may be chronic rather than acute, and this will introduce further uncertainty. The test may have been an LC 50 test, while the criteria standards may involve NOELs, which again have to be uncertainly estimated from the LC 50. Typically these uncertainties are combined into a single inferential step, often by assuming worst case in each step, and independence of each uncertainty. This procedure results in a conservative estimate, but rarely an accurate one. Further, it can create an unwarranted variance of several orders of magnitude from the actual test results. This type of inference procedure constitutes a probabilistic reasoning system, for which a number of mathematical formalisms have been developed in the artificial intelligence tradition, such as Dempster-Shafer theory, truth maintenance systems, and nonmonotonic logic. In this paper, we use several cases to illustrate the differences between the conventional approach and a more sophisticated approach that takes into account possible interactions between the various uncertainties in the system. It is generally possible to get much more realistic bounds on the risk assessment by invoking mathematical methods more sensitive to the logic of combined probabilities.

Keywords: uncertainty, risk assessment, probability, artificial intelligence, expert systems

> Life is the art of drawing sufficient conclusions from insufficient premises. —*Samuel Butler*

1 Introduction

Risk assessment involves the combination of a wide variety of more or less uncertain sources of information. Some are known very accurately, such as the gravitational constant or the balances required in redox equations, others are known approximately, such as the LC 50 of copper sulfate for rodents, while others are largely informed conjecture, such as the strength of a public reaction to a 10% increase in the acidity of rain or the stability of an ecosystem. Usually, each of these uncertainties is modelled by a *probability distribution over the possible values* that each of the variables or parameters of interest can obtain. We discuss here several approaches to uncertain reasoning that come out of the artificial intelligence (AI) tradition, and how use of these techniques might improve the practice of risk assessment.

The variables that go into a risk assessment can be grouped into three major categories:

1. Physical parameters.
2. Decisions.
3. Values.

Physical parameters are things like temperature, pH, number of organisms, and so on. In purely scientific studies, as opposed to policy making studies, physical pa-

rameters are often the only variables that go into the analysis. Decision parameters are items that are under the user's control. The decision to grant permits, for example, can take on such values as: no permits, a few restricted permits, or permits granted to all who apply. The values of the physical variables often feed into the decisions, but generally decisions are made in the hope of maximizing the value parameters. Value parameters are things like jobs, clean air, and healthy wildlife populations.

Establishing reasonable values for these uncertain quantities is a difficult enough task. However, even after the experiments or surveys have been done, the problem remains of *combining* various uncertain quantities, of reasoning from one unsure foundation to another. For example, one may have reasonably accurate information about the relation of a toxin to a particular species, and reasonably accurate information about the structure of the toxin and its toxic relationship to various metabolic pathways, but need to extrapolate this evidence to other species, to an entire ecosystem, or to other toxins. Methodologies such as the QSAR, for example, are attempts to extrapolate from tested species to untested species, *e.g.* rats to *Daphnia*, or from tested compounds to untested compounds, *e.g.* 2.4 dichlorophenol to 2.6 dichlorophenol (Enslein and Craig, 1978; Enslein et al., 1983; Enslein et al., 1988).

Typically, it is assumed that the uncertainties in an analysis are probabilities of one sort or another, and that, accordingly, the only appropriate models for combining them are the laws of probability. However, anylyzing a set of variables (including, perhaps, physical parameters, decisions, and values) with a mathematical, probabilistic model leads quickly to four major problems:

1. A combinatorial explosion of possibilities.

2. A lack of semantic information to guide inferences.

3. Poor methods of dealing with ignorance as well as uncertainty.

4. The need to calculate all values in the model at once, rather than incrementally as evidence is obtained.

Recent AI research has directly addressed these problems. In this paper we briefly consider some of the merits and problems of three AI approaches: localized approaches (which attempt to solve the combinatorial explosion problem), causal nets (which attempt to solve the semantic problem), and Dempster-Shafer calculus (which attempts to solve the ignorance problem). All of them have the benefit of being incremental approaches; as each new piece of information is added to the model, the model incorporates it without large-scale recomputation of all that has gone before.

After a brief introduction to the underlying probabilistic model of uncertainty analysis, we will discuss each of the three AI approaches in turn.

2 Mathematical model

The underlying probabilistic model is well understood in the risk assessment literature (Morgan and Henrion, 1990). If a problem concerns a set of variables, for example $\{A, B, C, D, E\}$, then, for each value that each variable can take on, we need to know the *joint probability* of that combination, $P(a, b, c, d, e)$ (where a is a value A can take on, *etc.*). The immediate problem with this approach is

that it is intractable for even small numbers of variables. If there are, say, only 20 variables in a problem, and each can take on, say, 6 values, then there are $6^{20} = 3,656,158,440,062,976$, over 3 quadrillion, different combinations of these values. Specifying all of these values is plainly unrealistic, but which values are necessary, and which redundant?

If the variables are continuous numbers and can, in effect, take on a infinite number of different values, then the joint probabilities must be specified as continuous multivariate functions of those variables, an even more daunting task. Generally speaking, most practical risk assessment proceeds by making all variables discrete; for example, species may be considered "highly susceptible," "moderately susceptible," or "not susceptible." To keep things simple, we will also, for the most part, assume that variables are categorical, that is, there are only a small number of discrete values they can take on. However, many of the techniques discussed can be generalized to the continuous case.

Characteristically, probabilities are not computed from a full, joint probability distribution, but are dealt with in a probability tree, such as the one in Figure 1. In this figure we have only four variables, and each variable (A, B, C, and D) has two possible values, which we will represent as $+a$, $-a$, *etc.*, and indicate by the upper and lower branches. There are, accordingly, $2^4 = 16$ possibilities, one for each path through the tree from left to right: the ends of the far-right arrows each represent a different possible outcome. The heavy arrows, for example, represent the combination $(+a, -b, -c, +d)$. The numbers on the arrows represent conditional probabilities, based on *all* the choices to the left. For instance, the heavy arrow above C in the figure has the value 0.8, indicating that the conditional

Figure 1 here.

probability of $-c$ given $+a$ and $-b$, is 0.8, written $P(-c|+a,-b) = 0.8$. If all 2^4 probabilities are known in advance (one number attached to each of the ends of the far-right arrows), then these conditional probabilities can be calculated by summing and dividing from right to left. The values at the top right, for example, indicating that $P(+a,+b,+c,+d) = 0.01$ and $P(+a,+b,+c,-d) = 0.004$ together imply that $P(+d|+a,+b,+c) = 0.01/(0.01 + 0.004)$, and so on. Likewise, knowing all of the conditional probabilities will determine the joint probabilities. The heavy arrows, for example, tell us that $P(+a,-b,-c,+d) = (0.3)(0.2)(0.8)(0.1) = 0.0048$.

It is usually much easier for humans to estimate a conditional probability than to estimate a joint probability. For instance, the probability that it rained last night, given that the grass is wet and you heard thunder, could be estimated. But estimating the probability that you will hear thunder tonight and find wet grass in the morning, unconditioned by anything, usually leads to confusion. Human probabilistic judgements are usually conditional, and therefore probability trees such as the one in Figure 1 are usually filled in along the branches, rather than from the right side.

The tree can, of course, be rearranged, putting B before A, etc., and getting a different set of conditional probabilities ($P(+a|-b)$ instead of $P(-b|+a)$, for instance). However, there are still an insuperably large number of conditional probabilities that must be estimated, and the mathematical model itself gives us no help in determining which are relevant and which irrelevant. Further, if there are some probabilities in the tree about which we are largely, or even completely, ignorant, *some* values for them will have to be provided, even if they are completely arbitrary. In situations of complete ignorance, a uniform probability distribution is usually as-

sumed: all outcomes equally likely. Other situations require a "seat-of-the-pants" estimate; for example, we may estimate that 75% of the local population is likely to favor a pesticide regulation, using only the current political climate as guidance. This is not total ignorance, but it is just as arbitrary.

These problems: huge numbers of possibilities, not knowing which of them are relevant, treating ignorance in an *ad hoc* manner, and the basic need to recalculate everything when any one thing changes, lead us into several models of reasoning under uncertainty that stem from the AI tradition. We now turn to a consideration of three of them, and their relative merits in dealing with these problems.

3 Local approaches

Early in the development of expert systems, the combinatorial problems associated with inference under uncertainty were recognized. While it was recognized that, if the presence of a was evidence for b (*e.g.* $P(b|a)$ was high), then even if we know a is true we still cannot conclude anything about b without knowing if a is the *only information relevant* to b. Another factor, such as c, might completely alter our expectations. For example, elevated temperature in an aquatic system generally connotes reduced dissolved oxygen concentrations because of the inverse relationship between oxygen solubility and temperature. However, the elevated temperature may also imply that it is mid-summer. Photosynthetic activity during this time may cause increased dissolved oxygen levels if the values come from the epilimnion of a biologically productive lake (see Figure 2).

Because it was clearly unrealistic for every inference to consult every possibly

relevant fact in the system, an approximate approach was used, which would go ahead and make inferences from a to b, but would attach "certainty factors" to the conclusions. Certainty factors are definitely *not* probabilities; calculating probabilities was deemed too hard and certainty factors were a substitute. An example from the MYCIN system follows (Buchanan and Shortliffe, 1984). MYCIN was an early expert system constructed to perform medical diagnosis: examine symptoms, recommend further tests, and make inferences as to likely causes.

Each inference rule in MYCIN was expressed as an "if-then" statement with a certainty factor attached, such as these:

1. If a then c (0.4)
2. If b then c (0.6)
3. If c then d (0.8)

which indicated that, for example, if you were reasonably sure about c, then you would be 80% as sure about d. Various combination rules had to be devised when chains of reasoning were involved. For example, if a and b were both known for certain, the first two rules could be combined under the following formula to get a certainty factor for c:

$$\begin{aligned} CF(c) &= 0.4 + 0.6 - (0.4)(0.6) \\ &= 0.76 \end{aligned}$$

Given this certainty factor for c, the third rule above could be used to give a certainty factor for d:

$$CF(d) = (0.76)(0.8)$$

$$= 0.61$$

The MYCIN certainty factors take on both positive and negative values, allowing evidence to be either for or against a conclusion.

Such localized rules essentially solved the combinatorial explosion problem by *ignoring it*. Their use resulted in practical, working systems that solved large problems in the real world (Buchanan and Shortliffe, 1984). However, they had to be used with great care, because, strictly speaking, their inferences were invalid. Consider, for example, what would happen with these rules if different *types* of reasoning are mixed. Some inferences are from cause to effect; for example, if you open the floodgates, you can safely infer that the water downstream will rise. On the other hand, some inferences are from effect to cause; for example, if you find a large fish kill, you can legitimately raise your expectation of toxins in the water. But putting two such inferences together can be disastrous. Consider:

- If the sprinkler was on then the grass is wet (0.9)
- If the grass is wet then it rained (0.8)

Therefore:

- If the sprinkler was on then it rained

  (0.9 * 0.8 = 0.72)

Each of the two original inferences is quite probable; each of their "if" parts lends support to their "then" parts. The combination of the two, however, is ludicrous.

One attempt to incorporate information such as cause-effect relationships into the process of reasoning under uncertainty is provided by causal nets, considered in the next section.

4 Causal nets

Causal nets, also called Bayesian networks or influence diagrams, are an attempt to retain the original probabilistic model, exemplified in Figure 1, but meet head-on the problem of combinatorial explosion by analyzing the *kinds* of links in the diagram, and reducing the number of calculations that have to be done without sacrificing validity of the inferences (Pearl, 1988).

One of the devices brought to bear on this problem is distinguishing cause and effect, as mentioned at the end of the last section. In Figure 3, the inferences from "sprinkler" to "grass" and from "grass" to "rain" are distinguished by being in the opposite causal direction. Inferences from cause to effect are carried by π-messages, while inferences from effect to cause are carried by λ-messages. (Since we normally have conditional probabilities of effects, given causes, π's are associated with probabilities while λ's are associated with likelihoods, hence the names.) Careful handling of λ and π messages at each point avoids the nonsensical inference from "sprinkler" to "rain", but does so in a way that does *not* require every inference to check every other fact in the system before going ahead. In fact, only in certain, restricted classes of systems does any non-local checking have to be done. Causal "loops" are one example, where, for instance, a single cause can have two effects, but each effect can result in the same symptom. In Figure 4, for instance, the observation of increased chlorophyll would naturally lead to an increased probability of algal enhancement, which should strengthen the probability of *both* an oxygen sag (by a π message) and the probability of some form of nutrient enhancement (by a λ message). However, the oxygen sag should not then send a λ message up the

Figur
here.

Figu
here

fish kill → nutrient ladder, because this would increase the probability of nutrients *twice* on the *same* piece of evidence.

Such loops raise problems for the causal net model, and there are a number of approaches to dealing with them; but these problems are minor compared to a straightforward mathematical model which would require *all* factors be reconsidered in *all* inferences.

A number of other advantages to the causal net model come about as well. The importance of *qualitative* uncertainties is obvious. The EPA Framework for Ecological Assessment, for example, asserts that,

> ... often the relationship [between measurement and assessment endpoints] can be described only qualitatively. Because of the lack of standard methods for many of these analyses, professional judgment is an essential component of the evaluation (U. S. Environmental Protection Agency, 1992, p. 23)

However, a causal net model offers a standard, formal, and qualitative treatment of independence. In the mathematical model, for example, independence of events is defined quantitatively, based on the probability distributions: a is said to be independent of b, given c, if and only if

$$P(a|b,c) = P(a|c)$$

Clearly, to establish this in general, one has to go back to the joint probabilities and calculate things numerically. Humans, however, can often judge whether two things are independent, without having the slightest idea of the numeric probabilities involved. Consider, for instance, a watershed study and the question of whether or

not rainfall is independent of soil type. Normally we could easily judge that these two factors are independent. However, to verify this mathematically, the joint probabilities for each plot of land, for each amount of rain, and for each soil type, would all have to be calculated or estimated. This is clearly a large task, and also plainly a waste of time given that we can judge their independence qualitatively without any of the numbers.

Causal nets, on the other hand, by distinguishing π (cause to effect) and λ (effect to cause) inferences, can give deep qualitative insight into this kind of independence. For example, height and reading ability in humans are highly correlated. However, if you know a subject's age (presumably the root cause of the correlation between height and reading ability), then height and reading ability become independent. On the other hand, earthquakes and burglaries are largely independent, but both can cause your car-alarm to go off. Hearing your car alarm simultaneously raises the probability of both a burglary and an earthquake, but also renders them dependent—hearing about an earthquake on your radio will decrease your expectation of a burglar at your car. Rainfall and soil type, for another example, are only *conditionally* independent. If it is learned that a hill slope failure occurred, then rainfall and soil type are no longer independent: a very stable soil type would increase the probability of heavy rain before the failure. Causal nets, in conjunction with algorithmic inference engines, can automate such complex qualitative reasoning. The automation of such inferences becomes critical as the systems dealt with become more complicated, and dozens or hundreds of intertwined causes and effects begin to interact.

An extension of the causal net model to continuous-valued numeric variables is

straightforward (Pearl, 1988, pp. 344-356), and only requires that some tractable model of the uncertainties be used. The usual assumptions about uncertainties, such as uncorrelated, normal distributions, and linear interactions between variables, suffice.

5 Dempster-Shafer theory

Causal nets are an improved reasoning tool for dealing with probabilities such as those found in the standard model (Figure 1). However, even with the improvements found in a causal net approach, at times the probabilities in the standard model remain intractable. Dempster-Shafer theory was designed to overcome some of these problems, by approaching probabilities in an entirely different light (Shafer, 1976; Gordon and Shortliffe, 1984). To understand this approach, consider a standard model with just two variables, a and b. In the standard model, probabilities must be assigned to all possible outcomes, namely, $(+a, +b)$, $(+a, -b)$, $(-a, +b)$, and $(-a, -b)$. Even in a situation of total ignorance, *some* probabilities (such as 0.25 to each) would have to be assigned to these. In the Dempster-Shafer model, *sets* of possible outcomes are considered. Probabilities are defined over these sets, denoting the hypothesis, in each case, that one or another of the possible outcomes in the set will be the true one. In our two variable example, for instance, the sets might consist of such things as $\{(+a, +b), (-a, -b)\}$, denoting the hypothesis that either *both* a and b will be the case, or neither will, or $\{(+a, -b), (-a, +b)\}$, denoting the hypothesis that if either a or b happens, the other won't.

The logic of this approach thus contrasts with the standard model. Rather

than making joint probabilities easier to deal with by breaking them down into conditional probabilities, joint probabilities are simplified by lumping them together. The intuition is that many working hypotheses in science are of this nature: a disease symptom, for example, may indicate one of several diseases and eliminate others. The presence of such a symptom, then, is evidence for an hypothesis that is essentially a disjunction: it's probably either A or B or C, where each of the hypotheses (A, B, and C) is itself a *complete* specification of the system.

This approach has the advantage of immediately simplifying most problems. In dealing with a complex ecological system, for instance, a natural approach does not usually involve hypotheses governing all possible states of all variables in all combinations. Rather, a few models are conjectured that have consequences for *all* of the variables. For example, a eutrophic lake would characteristically imply high temperature, low dissolved oxygen, and a deep depth. An oligotrophic lake, on the other hand, would imply high temperature, high dissolved oxygen, and either deep or shallow depth. More finely divided scenarios would be devised, of course, to fit the level of assessment desired.

Further, the calculation of probabilities over these sets is freed from some of the problems that plague causal nets and other "Bayesian" approaches. The selection of prior probabilities, for example, is eliminated. Rather than, say, assigning a uniform probability to all possible outcomes in the case of complete ignorance, the Dempster-Shafer theorist simply assigns probability one to the set of all possible outcomes (a set usually denoted by Θ, and called the *frame of discernment*), and zero to any subset. To make sure these probabilities of *sets* of hypotheses are not confused with probabilities of hypotheses, we use m instead of P, and say $m(\Theta) = 1.0$. In

a Bayesian approach, by contrast, the initial state of ignorance might be modelled using a uniform distribution; for example, if there were n possible outcomes, each one would be assigned a probability of $1/n$.

For a simple example of subsequent calculations and the incremental propagation of uncertain information in the Dempster-Shafer model, consider a simple situation in which there are only three possible outcomes, A, B, and C. All possible subsets of these outcomes are illustrated in Figure 5 (except the empty set, which, by assumption, will never have a probability greater than 0). The frame of discernment $\Theta = \{A, B, C\}$ is at the top, and the subset relation is indicated by an arrow. Initially,

Fig her

$$\begin{aligned} m(\Theta) &= 1.0 \\ m(\{A, B\}) &= m(\{A, C\}) = \ldots = 0.0 \end{aligned}$$

(A Bayesian approach, on the other hand, would have $P(A) = P(B) = P(C) = 1/3$.) Now suppose that information is gained suggesting, at a level of 0.6, that either B or C is correct. We update as:

$$\begin{aligned} m(\Theta) &= 0.4 \\ m(\{B, C\}) &= 0.6 \\ m(\{A, B\}) &= m(\{A, C\}) = \ldots = 0.0 \end{aligned}$$

Notice that the remainder ($0.4 = 1.0 - 0.6$) is *not* assigned to $\{A\}$, the *complement* of $\{B, C\}$, but remains with the completely neutral hypothesis set, $\{A, B, C\}$. This accords well with intuitions: evidence in favor of $\{B, C\}$ should not *increase* the probability of $\{A\}$ from 0 to 0.4.

Combining further evidence with this m function proceeds as follows. Let us call the above function m_1, and suppose we gain evidence in favor of $\{A, B\}$, with strength 0.5. This would give us a new function, m_2, with

$$\begin{aligned} m_2(\Theta) &= 0.5 \\ m_2(\{A, B\}) &= 0.5 \\ m_2(\{B, C\}) &= m_2(\{A, C\}) = \ldots = 0.0 \end{aligned}$$

In this case, we would expect B to be supported at some level greater than zero, since it was supported by both pieces of evidence, and this is the case. The combined measure function, m_3, obtained from m_1 and m_2, is defined as follows, for any set Z:

$$m_3(Z) = \sum_{X \cap Y = Z} m_1(X) \cdot m_2(Y)$$

Accordingly,

$$\begin{aligned} m_3(\{B\}) &= m_1(\{B, C\}) \cdot m_2(\{A, B\}) \\ &= (0.6)(0.5) \\ &= 0.3 \\ m_3(\{A, B\}) &= m_1(\{A, B, C\}) \cdot m_2(\{A, B\}) \\ &= (0.4)(0.5) \\ &= 0.2 \\ m_3(\{B, C\}) &= m_1(\{B, C\}) \cdot m_2(\{A, B, C\}) \\ &= (0.6)(0.5) \\ &= 0.3 \end{aligned}$$

$$\begin{aligned} m_3(\{A,B,C\}) &= m_1(\{A,B,C\}) \cdot m_2(\{A,B,C\}) \\ &= (0.4)(0.5) \\ &= 0.2 \end{aligned}$$

and all other m_3 values are zero. Notice that the sum of all m_3 values remains one, as a probability distribution should. Occasionally, when evidence supports mutually incompatable hypotheses, the sum drops below one. For example, if one experiment supported A as the only explanation, and another experiment supported only B, then the empty set, $\emptyset = \{A\} \cap \{B\}$, representing "no possible explanation of the evidence," would get some amount of support. In this case, Dempster-Shafer theory specifies that the probabilities of the nonempty sets are simply scaled up so that the total sum remains one. Thus, the full equation for m_3, given m_1 and m_2, is:

$$m_3(Z) = \frac{\sum_{X \cap Y = Z} m_1(X) \cdot m_2(Y)}{1 - \sum_{X \cap Y = \emptyset} m_1(X) \cdot m_2(Y)}$$

This equation can be applied in an incremental fashion as each piece of information is acquired, or each decision contemplated.

These calculations may appear confusing and involved, and their justification involves deep results in model theory and logic (Shafer, 1976), but they are nonetheless intuitively satisfying and they can be fully automated. The important fact to notice about them is that practitioners, in dealing with uncertain evidence, need only specify which *sets of hypotheses* the evidence supports. The precise impact of a piece of evidence on any one variable, physical parameter, decision, or value, need not be estimated. Combinations of particular variables can be combined into scenarios, and the probabilities of each scenario dealt with directly. This can result in considerable conceptual clarity in dealing with complex situations. The usual

requirements of expert solicitation, that he or she imagine wildly unlikely combinations of events, and then estimate probabilities for other variables conditioned on them, are absent from the Dempster-Shafer methodology. Only likely scenarios, combinations of variable values, need be considered.

6 Conclusion

The logic of combined probabilities, studied extensively in the artificial intelligence tradition, is amenable to a large number of approaches. The mathematical foundations of probability are usually based on building up definitions and theorems based on complete knowledge of a joint probability distribution. However, the higher-level reasoning often pursued by humans in their assessment of uncertainty and risk often has little or no basis in numerical combinations of a huge number of probability estimates. Nevertheless, current practice in risk assessment often assumes that such rock-bottom numbers must be obtained or estimated, by some means, before uncertain inference can proceed.

We have outlined three recent approaches to uncertain inference that stem from the artificial intelligence tradition. Localizing the inferences allows us to forget about many of the numbers involved, but at the expense of making quite unreliable inferences at times. Causal nets reduce some of the complexity of the problem, can support automated qualitative reasoning about uncertainty, and are faithful to the cause/effect distinction which permeates uncertain reasoning. Dempster-Shafer theory allows uncertain reasoning to proceed on a different level, on the level of sets of likely scenarios rather than sets of variables and their values, and as a result

greatly reduces the effort in translating human intuition into an automated system, and has a much more intuitively satisfying treatment of ignorance.

The ability to automate each of these approaches, to embody their inference structure into a computer program, has the potential for even greater rewards. A long tradition of machine learning has found that often a computer-generated analysis can be superior to human intuition. A strong example is provided by Michalski's expert system (Michalski and Chilausky, 1980). Michalski and his colleagues went through a long consultation phase with a human expert in soybean pathology in an effort to build an expert system capable of diagnosing soybean diseases. Michalski then used a machine learning system to build a second expert system solely *from data* concerning soybean diseases and their symptoms; in other words, he used another AI program, a *learning* program, to extract the rules used by the second expert system. Both expert systems were then tested on new cases. The set of rules produced by the human pathologist correctly identified only 83% of the new diseases, while the set of rules produced by the computer program correctly identified 99.5% of the new cases. "...plant pathologists are now using the machine-induced rules for their routine diagnoses" (Firebaugh, 1988).

A recent study of the future of computer science and engineering (CS&E) by a committee of the National Research Council concluded that recent advances in CS&E were not readily available to many other disciplines, and called on CS&E to increase its interactions with other disciplines. Among the top priorities for the future of CS&E they listed:

- Increase its contact and intellectual interchange with other disciplines ...
- Increase the number of applications of computing and the quality of existing applications in areas of economic, commercial, and social significance ...
- Increase traffic in CS&E-related knowledge and problems among academia, industry, and society at large, and enhance the cross-fertilization of ideas in CS&E between theoretical underpinnings and experimental experience

(Committee to Assess the Scope and Direction of Computer Science and Technology, NRC, 1992, p. 34)

This paper is an attempt to initiate a dialogue between CS&E professionals versed in many techniques of automated reasoning under uncertainty and the practitioners of risk assessment nationwide. Each of the approaches sketched here has great potential in risk assessment, particularly in automated software tools which may soon form a critical part of the risk analyst's repertoire.

7 Acknowledgements

This research is supported by USAFOSR Grant No. AFOSR-91-0291 DEF.

References

Buchanan, B. G. and Shortliffe, E. H. (1984). *Rule-Based Expert Systems, The*

MYCIN Experiments of the Stanford Heuristic Programming Project. Addison-Wesley, Reading, Massachusetts.

Committee to Assess the Scope and Direction of Computer Science and Technology, NRC (1992). Computing the future. *Communications of the ACM*, 35(11):30–40.

Enslein, K., Blake, B. W., Tuzzeo, T. M., Borgsdtedt, H. H., Hart, J. B., and Salem, H. (1988). Estimation of rabbit eye irritation scores by structure-activity equations. *In Vitro Toxicology*, 2(1):1–14.

Enslein, K. and Craig, P. N. (1978). A toxicity estimation model. *Journal of Environmental Pathology and Toxicology*, 2:115–121.

Enslein, K., Lander, T. R., Tomb, M. E., and Landis, W. G. (1983). Mutagenicity (ames): A structure-activity model. *Teratogenesis, Carcinogenesis, and Mutagenesis*, 3:503–513.

Firebaugh, M. W. (1988). *Artificial Intelligence: A Knowledge-Based Approach.* Boyd & Fraser Publishing Company, Boston, MA.

Gordon, J. and Shortliffe, E. H. (1984). The Dempster-Shafer theory of inference. In Buchanan, B. G. and Shortliffe, E. H., editors, *Rule-Based Expert Systems, The MYCIN Experiments of the Stanford Heuristic Programming Project*, pages 272–292. Addison-Wesley, Reading, Massachusetts.

Michalski, R. S. and Chilausky, R. L. (1980). Learning by being told and learning from examples: an experimental comparison of the two methods of knowledge

acquisition in the context of developing an expert system for soybean diseases. *Policy Analysis and Information Systems*. 4.

Morgan, M. G. and Henrion, M. (1990). *Uncertainty, A Guide to Dealing with Uncertinty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, California.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ.

U. S. Environmental Protection Agency (1992). *Framework for Ecological Risk Assessment*. EPA/630/R-92/001.

Legends for Figures

Figure 1. Basic probability model. Each path from left to right represents a combination of the variables A, B, C, and D. Conditional probabilities lie along arrows, joint probabilities are found at the extreme right hand side.

Figure 2. A case in which one cause (high temperature) can lead to different effects in different circumstances. The conditional probability alone of low dissolved oxygen, given high temperature, does not allow an inference from high temperature to low dissolved oxygen.

Figure 3. Bayesian inference takes account of cause and effect by distinguishing inferences based on causes (π inferences) from inferences based on effects (λ inferences).

Figure 4. A causal loop that must be handled carefully in Bayesian inference, even if π anf λ inferences are distinguished.

Figure 5. Dempster-Shafer theory calculates probability over sets of hypotheses, not single variable values. This illustration shows all possible subsets of three hypotheses.

Figures

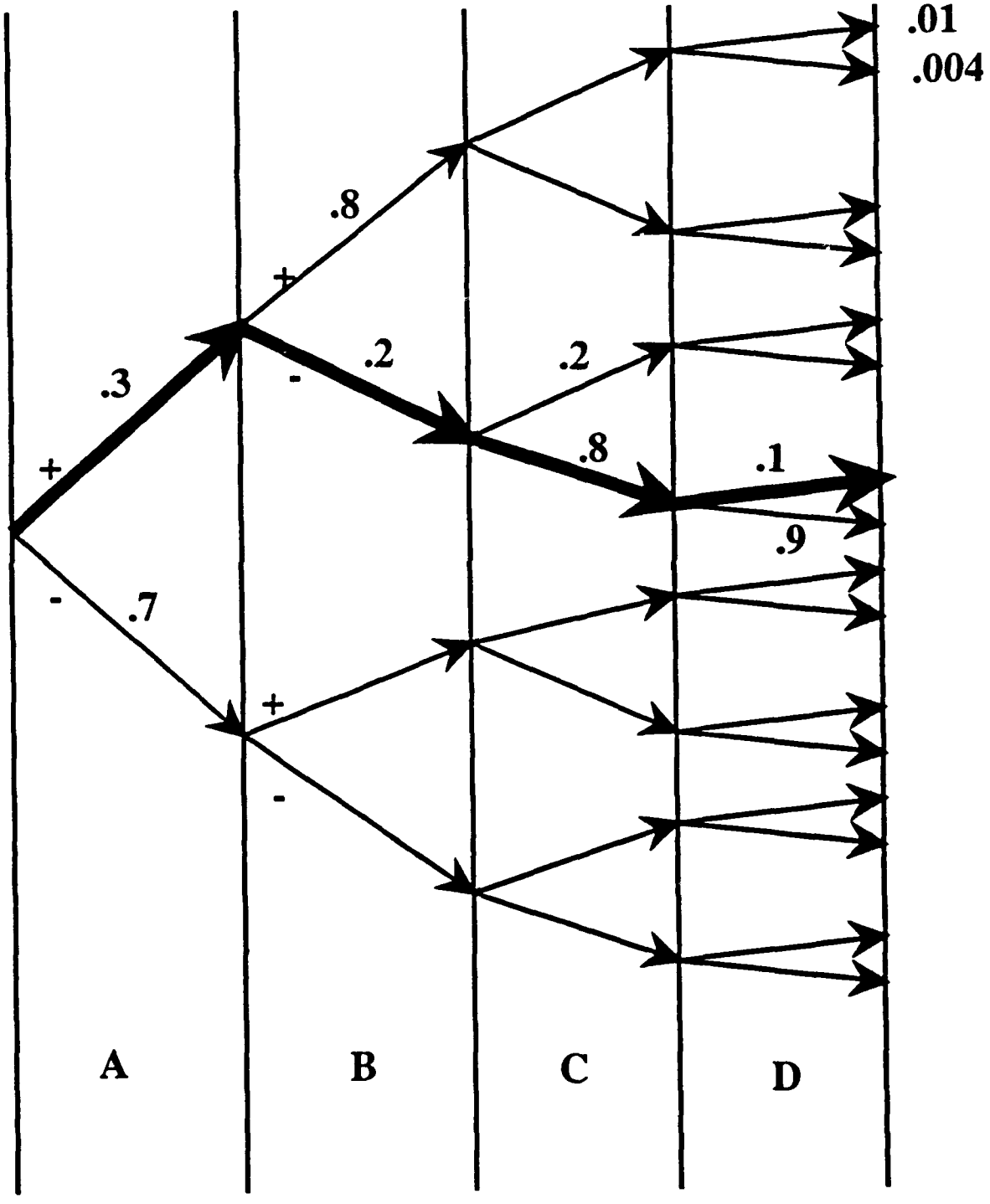


Figure 1

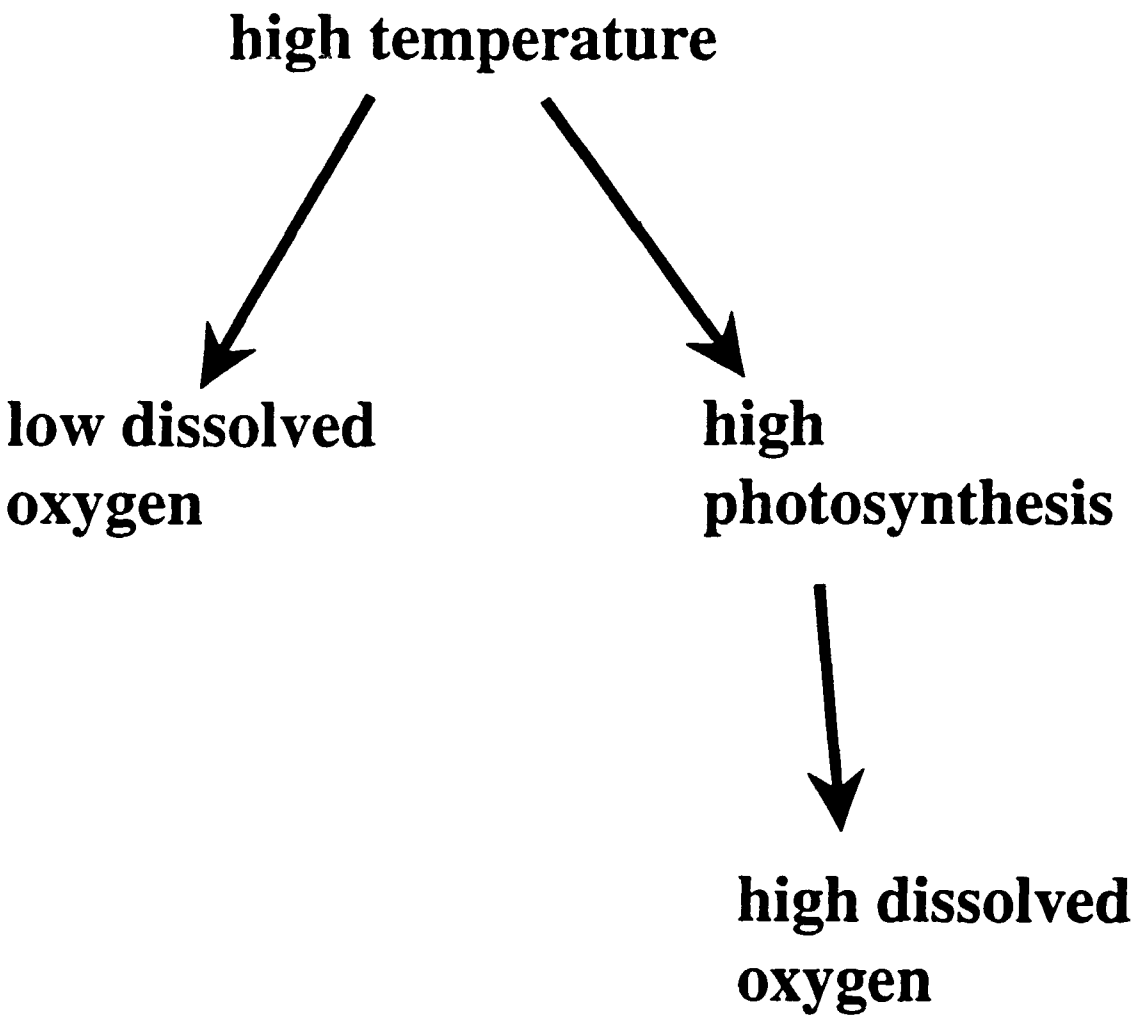

high temperature
low dissolved oxygen
high photosynthesis
high dissolved oxygen

Figure 2

Bayesian Networks

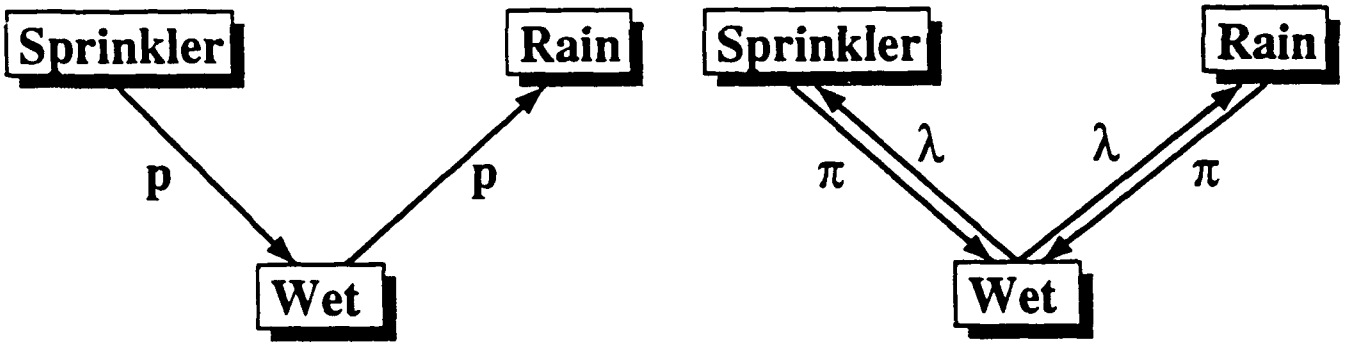


Figure 3

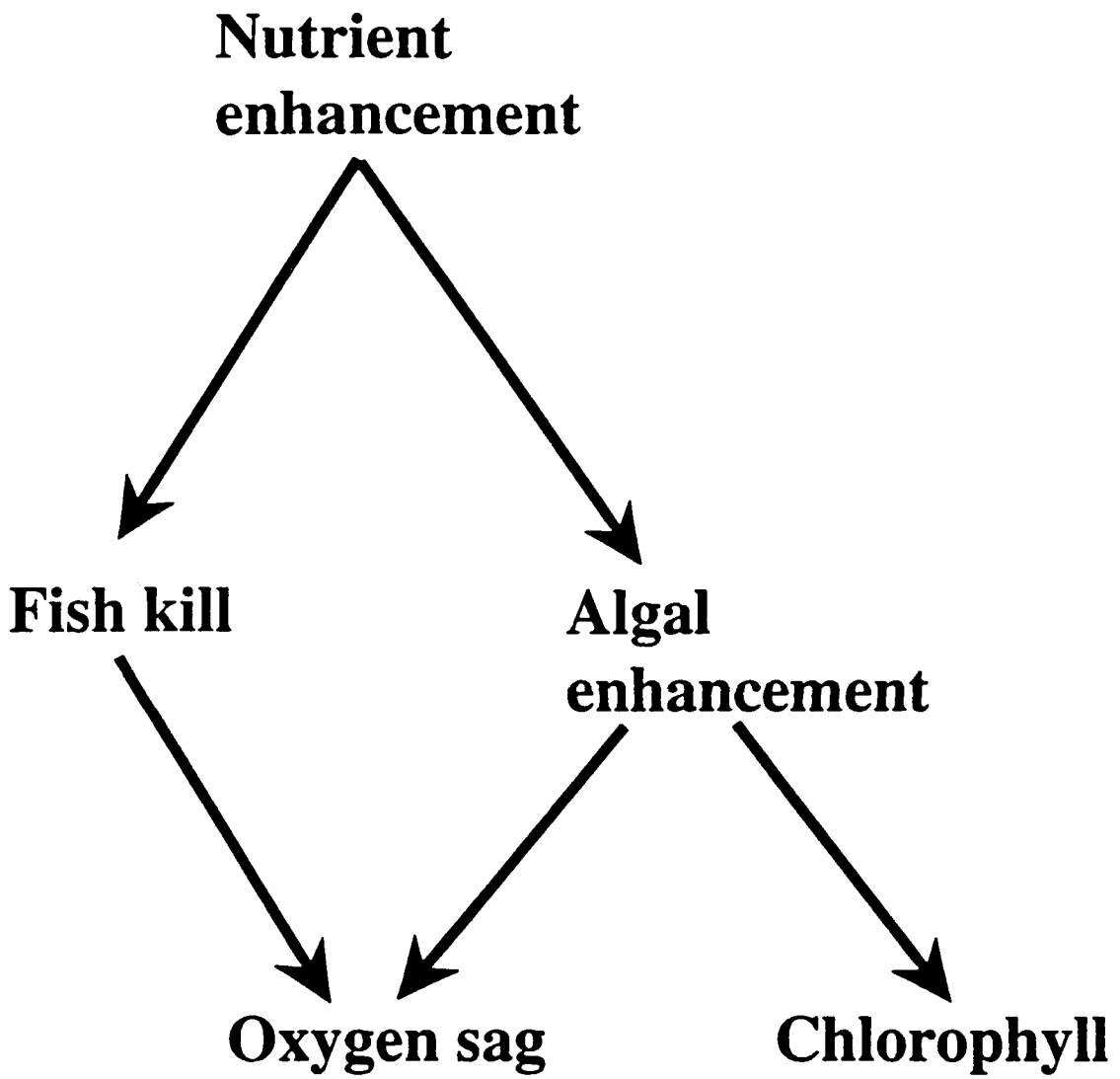

Nutrient enhancement
Fish kill
Algal enhancement
Oxygen sag
Chlorophyll

Figure 4

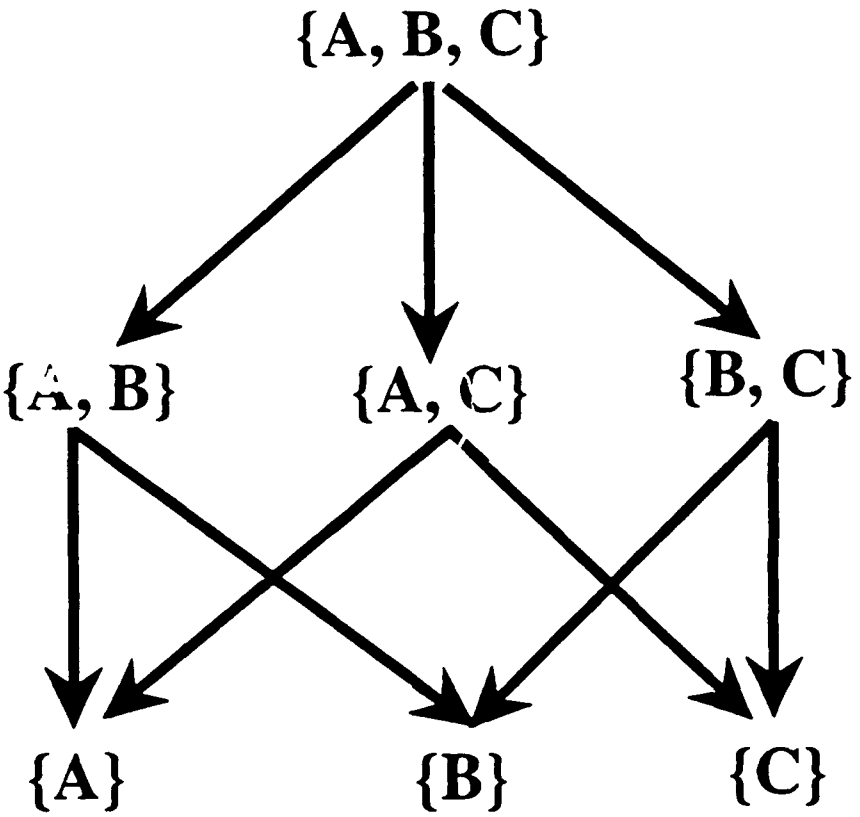


Figure 5

Submitted ASTM 1993

1)Draft Report, Please do not cite or quote

1

Wayne G. Landis[1], Robin A. Matthews[1], April J. Markiewicz[1] and Geoffery B. Matthews[2].

Non-linear Oscillations Detected By Multivariate Analysis in Microcosm Toxicity Tests with Complex Toxicants: Implications for Biomonitoring and Risk Assessment.

REFERENCE: Landis, W. G., Matthews, R. A., Markiewicz, A. J. and Matthews, G. B. **"Non-linear Oscillations Detected By Multivariate Analysis in Microcosm Toxicity Tests with Complex Toxicants: Implications for Biomonitoring and Risk Assessment,"** Environmental Toxicology and Risk Assessment-Third Volume, ASTM 1218, Jane S. Hughes, Gregory R. Biddinger, and Eugene Mones, Eds., American Society for Testing and Materials, Philadelphia, 1994.

A common assumption in environmental toxicology is that after the initial stress, ecosystems recover to resemble the control state. This assumption may be based more on our inability to observe an ecosystem with sufficient resolution to detect differences, than reality. This study compares the dynamics of the effects of the water soluble fraction (WSF) of both Jet-A and JP-4 using the Standard Aquatic Microcosm (SAM) using several types of multivariate analysis.

Two SAM experiments have been completed using concentrations of 0.0, 1, 5 and 15 percent WSF. The effects of the WSF on the microcosm communities were subtle. Among the more interesting effects were the shifts in time of population peaks and some other variables compared to reference microcosms. In both experiments, multivariate analysis was able to differentiate oscillations that separate the treatments from the reference group, followed by what would normally appear as recovery, followed by another separation into treatment groups as distinct from the reference treatment. These patterns generally were not detected by conventional analysis.

Two sets of related explanations exist for the observed phenomenon. First, the addition of the toxicant initiates an alteration in the community so that the quality of the food resources for the later successional stages is significantly different from the control. This difference in resource quality and quantity leads to the repeated and replicated oscillations. The second explanation is that the oscillations are the result of the intrinsic chaotic behavior of population interactions, of which the alteration of detrital quality is but one of many. The initial impact of the toxicant re-set the dosed communities into different regions of the n-dimensional space where recovery may be an illusion due to the incidental overlap of the oscillation trajectories occurring along a few axes. Some of the implications of non-linear or chaotic dynamics upon the prediction of ecological risk are discussed.

Key Words: Standardized Aquatic Microcosm, jet fuel, non-linear dynamics, nonmetric clustering and association analysis, risk assessment

[1]Institute of Environmental Toxicology and Chemistry, Huxley College of Environmental Studies and
[2]Computer Science Department, Western Washington University, Bellingham, WA 98225.

INTRODUCTION

Over the last 15 years a variety of multispecies toxicity tests have been developed with the hope that in doing so, the increased complexity of the test would result in a more realistic comparison to community-level responses to the toxicant. However, the addition of more than one species, and the generally longer time periods associated with these multispecies tests, also result in much more complex data sets. Distinguishing toxicant effects from other community-level changes has become one of the most critical obstacles to the interpretation of multispecies data sets.

Multispecies toxicity tests are usually referred to as microcosms or mesocosms, although a clear definition of the size or complexity to distinguish these terms has not been put forth. In the Standardized Aquatic Microcosm (SAM) developed by Taub and colleagues (Taub 1969, 1976, 1988, 1989, Taub and Crow 1978, Crow and Taub 1979, Taub et al. 1980, 1987, 1988, Kindig et al. 1983, Conquest and Taub 1989) the physical, chemical, and biological components are defined as to species, media and substrate. The SAM system has undergone round robin testing (Conquest and Taub 1989) and has been used with a variety of toxicants and degradative organisms (Landis et al. 1989, 1993).

One of the major difficulties in the evaluation of multispecies toxicity tests has been the difficulty in the analysis of the large data set on a level consistent with the goals of the toxicity test. Typically, the goals of the multispecies toxicity test are twofold:

- to detect changes in the population dynamics of the individual taxa that would not be apparent in single species tests; and,

- to detect community-level differences that are correlated with treatment groups thereby representing a deviation from the control group.

A number of methods have been developed in an attempt to satisfy the goals of multispecies toxicity testing. Analysis of variance (ANOVA) is the classical method to examine single variable differences from the control group. However, because multispecies toxicity tests generally run for weeks or even months, there are problems with using conventional ANOVA. These include the increasing likelihood of introducing a Type II error (accepting a false null-hypothesis), temporal dependence of the variables, and the difficulty of graphically representing the data set. Conquest and Taub (1989) developed a method to overcome some of the problems by using intervals of non-significant difference (IND). This method corrects for the likelihood of Type II errors and produces intervals that are easily graphed, facilitating further analysis. The method is routinely used to examine data from SAM toxicity tests, and it is applicable to other multivariate toxicity tests. The major drawback of the IND is the limitation of examining one variable at a time over the course of the experiment. While this method addresses the first goal in multispecies toxicity testing, listed above, it ignores the second. In many instances, community-level responses are not as straightforward as the classical predator/prey or nutrient limitation dynamics, that are usually selected as examples of single-species responses representing complex interactions.

Multivariate methods have proved promising as a method of incorporating all of the dimensions of an ecosystem. One of the first methods used in toxicity testing was the calculation of ecosystem strain developed by Kersting (1984, 1985, 1988) for a three compartment

microcosm. This method has the advantage of using all of the measured parameters of an ecosystem to look for treatment-related differences. At about the same time, Johnson (1988a, 1988b) developed a multivariate algorithm using the n-dimensional coordinates of a multivariate data set and the distances between these coordinates as a measure of divergence between treatment groups. Both of these methods have the advantage of examining the ecosystem as a whole rather than by single variables, and can track such processes as succession, recovery and the deviation of a system due to an anthropogenic input.

However, a major disadvantage of both these methods, and of many conventional multivariate methods, is that all of the data are often incorporated without regard to the units of measurement, or to the appropriateness of including all variables in the analysis. Random variables indiscriminately incorporated into the analysis, may contribute so much noise that they overshadow variables that do show treatment-related effects.

Ideally, a multivariate statistical test used for evaluating complex data sets will have the following characteristics:

- It will not combine counts from dissimilar taxa or other variable classifications by means of sums of squares, or other *ad hoc* mathematical techniques.

- It will not require transformations of the data.

- It will work without modification on incomplete data sets.

- It will work without further assumptions on different data types.

- Significance of a variable to the analysis will not be dependent on the absolute size of its count, so that taxa having a small total variance, i.e. rare taxa, can compete in importance with common taxa, and taxa with a large, random variance will not automatically be selected, to the exclusion of others.

- It will provide an integral measure of the quality of the analysis, i.e. whether the data set differs from a random collection of points.

- It will, in some cases, identify a subset of the variables that serve as reliable indicators of the physical and biological environment.

Recently developed for the analysis of ecological data, nonmetric clustering is a multivariate derivative of artificial intelligence research, that satisfies all these criteria and has the potential of circumventing many of the problems of conventional multivariate analysis.

In this paper, we use three multivariate techniques to compare patterns in the data sets from two SAM toxicity tests using turbine fuels. The multivariate techniques include two conventional tests based on the ratio of multivariate metric distances (Euclidean distance and cosine of the vector distance), and one relatively new program, RIFFLE, which employs nonmetric clustering and association analysis (Matthews and Hearne 1991). All three of the multivariate techniques have proven useful in analyzing complex ecological data sets (Matthews et al. 1991a, 1991b). Of the three, only nonmetric clustering meets all of the criteria listed above (Matthews and Matthews 1991).

EXPERIMENTAL METHOD

Reagents

All chemicals used in the culture of the organisms and in the formulation of the microcosm media were reagent grade or as specified by the ASTM method.

Jet-A was provided by Fliteline Services of Bellingham, Washington and was refined by Chevron. The sample was obtained from the sample valve used for quality control. The shipment lot was recorded and is on file. JP-4 was supplied by the U. S. Air Force Toxicology Laboratory at Wright Patterson, AFB, Ohio.

Water Soluble Fractions

The water soluble fraction was prepared in glassware washed in nonphosphate soap, rinsed, then soaked in 2N HCl for at least one hour, rinsed ten times with distilled water, dried and finally autoclaved for 30 minutes. Microcosm medium, T82MV, acted as the diluent for the water fraction of the WSF.

Twenty five mL of fuel is added to the two liter separatory funnel, and is agitated as follows: [1] shake separatory funnel for five minutes, releasing built up pressure as necessary; [2] allow funnel contents to remain undisturbed for 15 minutes; [3] shake contents for five minutes, allow to stand 15 minutes; [4] continue same pattern for a total time of one hour; and finally [5] allow separatory funnel contents to remain undisturbed for eight hours. At the end of this procedure the mixture was allowed to stand overnight. The next day all but 100 mL of T82MV/water soluble fraction of jet fuel mixture from the separatory funnel (leaving the lighter, insoluble fuel mixture in the flask) was drained into a cleaned, sterile 1 liter amber glass bottle and capped with a Teflon-lined screw cap. The WSF was used within 24 hours or stored at 4°C for no longer than 48 hours before use as the toxicant mixture.

Gas Chromatography of WSF

This protocol utilizes a Tekmar LSC 2000 Purge and Trap (P&T) concentrator system in tandem with a Hewlett Packard 5890A Gas Chromatograph with a Flame Ionization Detector (FID)(ASTM D3710, D2887, Westendorf 1986). Instrument blanks and deionized distilled water blanks are used to verify the P&T and GC columns cleanliness prior to analysis of samples. A five mL sample is injected into a five milliliter sparger, purged with pre-purified nitrogen gas for eleven minutes and dry purged for four minutes. Volatile hydrocarbons, purged from the sample and collected on the Tenax/Silica Gel column, are desorbed at 180°C directly onto the gas chromatograph SPB-5, 30m x 0.53 mm ID 1.5µm film, fused silica capillary column. The column, at 35°C, is held at that temperature for two minutes, increased to 225°C at 12°C/min and held at that temperature for five minutes. A Spectra-Physics 4290 Integrator records the FID signal output of the volatile hydrocarbons that have been separated and eluted from the column by molecular weight. A comparison is then made of the sample chromatograph to n-paraffin and n-naphtha chromatograph standards for sample concentration determinations.

Identification and Quantification of GC Fractions

Qualitative identification of some components in the WSF were determined using a Simulated Distillation (SIMDIS) Calibration Mixture. The ASTM Method D3710 Qualitative Calibration Mixture is the standard

test method for determining the Boiling Range Distribution of Gasoline and Gasoline Fractions by Gas Chromatography. This mixture was used as a calibration standard to determine the retention times for each known component in the mixture against which unknown components, in the WSF of the fuel mixture, were compared and identified.

SAM Protocol

The 64-day SAM-protocol previously has been described (ASTM E1366). Briefly, the microcosms were prepared by the introduction of ten algal, four invertebrate, and one bacterial species into 3L of sterile defined medium. Test containers were 4 L glass jars. An artificial sediment consisting of 200 g acid washed silica sand, cellulose and 0.5 g of ground chitin is autoclaved in the experimental jar; immersed in a water bath to a point above the level of the sediment during sterilization to prevent breakage.

Numbers of organisms, dissolved oxygen (DO) and pH were determined twice weekly. Room temperature was 20°C ± 2°. Illumination was 80.0 μEm^{-2} sec^{-1} PhAR with a range of 78.6-80.4 and a 12/12 day/night cycle.

Two major modifications were made to the SAM protocol. The first was the means of toxicant delivery. Test material was added on day 7 by stirring each microcosm, removing 450 mL from each container and then adding appropriate amounts of the WSF to produce concentrations of 0, 1, 5 and 15 percent WSF. After toxicant addition, the final volume was adjusted to 3L. No attempt to filter and retain the organisms withdrawn during the removal of the 450 mL was made prior to toxicant addition. All graphs and statistical analysis start with the next sampling day, day 11. The second modification was the substitution, in the JP-4 experiment, of *Tetrahymena thermophila* BIV for the hypotrichous ciliate. The hypotrichous ciliate was becoming increasingly difficult to culture, very likely due to the age of the clone. The results of the JP-4 study demonstrated the suitability of the Tetrahymena for inclusion in the protocol.

Data Analysis

All data were recorded onto standard computer entry forms and checked for accuracy. Parameters calculated included the concentrations of each of the species, DO, DO gain and loss, net photosynthesis/respiration ratio (P/R), pH, algal species diversity, algal biovolume, and biovolume of available algae. The statistical significance of these parameters, compared to the controls, was also computed for each sampling day using the IND plots developed by Conquest. The net photosynthesis/respiration ratio is not derived using ^{14}C methods but by comparing oxygen concentrations before lights on, at the end of the photosynthetic period just before lights off, and then at the next morning, as specified in the standard protocol. The photosynthesis/respiration ratio was then determined by incorporating these measurements.

The multivariate methods used in the analysis include cosine and vector distances and nonmetric clustering. All of these methods have been previously described (Matthews et al. 1991b, Landis et al. 1993) and are reviewed in this volume. Variables used in the multivariate analysis are presented in Table 1.

RESULTS

Persistence of the fuels. In the case of both WSFs, within three weeks after dosing the original material had been volitilized or degraded. In the case of JP-4, benzene, 2,4 dimethylpentane, ethylbenzene, 2-methylpentane, 2-methylpropane, o-xylene and toluene,

TABLE 1. Biotic parameters used in the multivariate statistical tests. Biotic variables such as diversity, available biovolume, and total algal biovolume are not used since they are derived from and therefore not independent of the variables listed below.

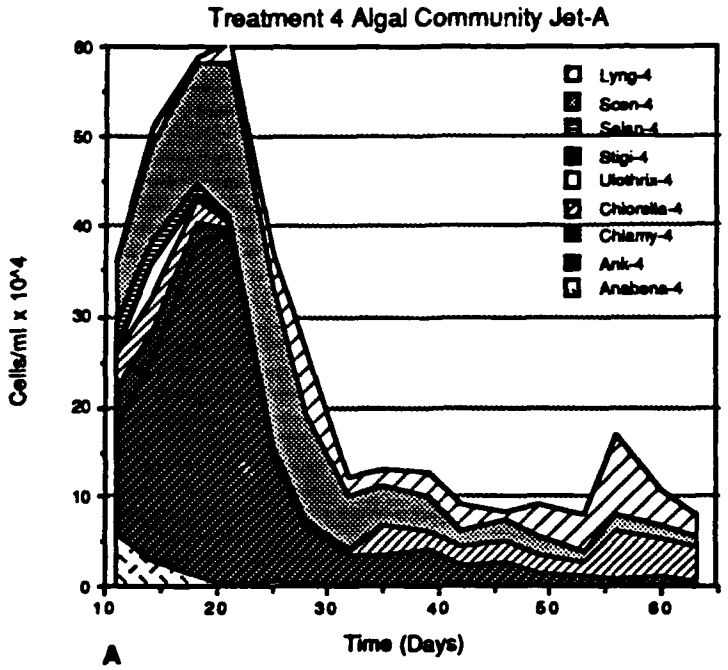
Jet A	JP4
Anabaena	Anabaena
Ankistrodesmus	Ankistrodesmus
Chlamydomonas	Chlamydomonas
Chlorella	Chlorella
Daphnia	Daphnia
Ephipia	Ephipia
Small Daphnia	Small Daphnia
Medium Daphnia	Medium Daphnia
Large Daphnia	Large Daphnia
Hypotricha	Tetrahymena
Lyngbya	Lyngbya
Miscellaneous sp.	Miscellaneous sp.
Ostracod (Cyprinotus)	Ostracod (Cyprinotus)
Philodina (Rotifer)	Philodina (Rotifer)
Scenedesmus	Scenedesmus
Selanastrum	Selanastrum
Stigeoclonium	Stigeoclonium
Ulothrix	Ulothrix

were tracked using GC analysis during the course of the SAM experiment. After week three, only 2-methylpentane and 2-methylpropane are detectable. Since only the 2-methylpropane is present 672 hours after dosing, this material may be the final biodegradative product of the absorbed fraction of the WSF, and is being investigated in more detail.

Comparison of Algal Population Dynamics-Highest Treatment. These area graphs (Figure 1) show the contribution of each algal species to the algal assemblage for the highest treatment concentration for each experiment. In the Jet-A treatment the algal populations were highest, reflecting the increased toxicity of the Jet-A to the daphnid populations. In both experiments however, an algal bloom was observed during the first 30 days of the experiment. At the end of the experiment the numbers and composition of the algal assemblage were similar, although the proportions of the species making up the assemblage had some differences. Chlorella seemed to be a greater constituent of the community in the JP-4 experiment.

Daphnid Population Dynamics. The most direct effect of the jet fuel upon the population dynamics of the daphnid populations was the delay in daphnid reproduction (Fig. 2). Peaks were delayed in the Treatment 4 microcosms in both instances. Daphnids were very important in determining the clusters in the early part of each experiment but not as important later. In both experiments two peaks of daphnid populations are observed. The first reflects the presence of the toxicant, the second occurs similarly in the dosed and not dosed systems. Error bars are not shown for clarity.

Ostracod Population Dynamics. Ostracod populations did not increase until late in each experiment (Fig. 3). In the Jet-A experiment (A), the numbers started an increase between days 40 and 45.



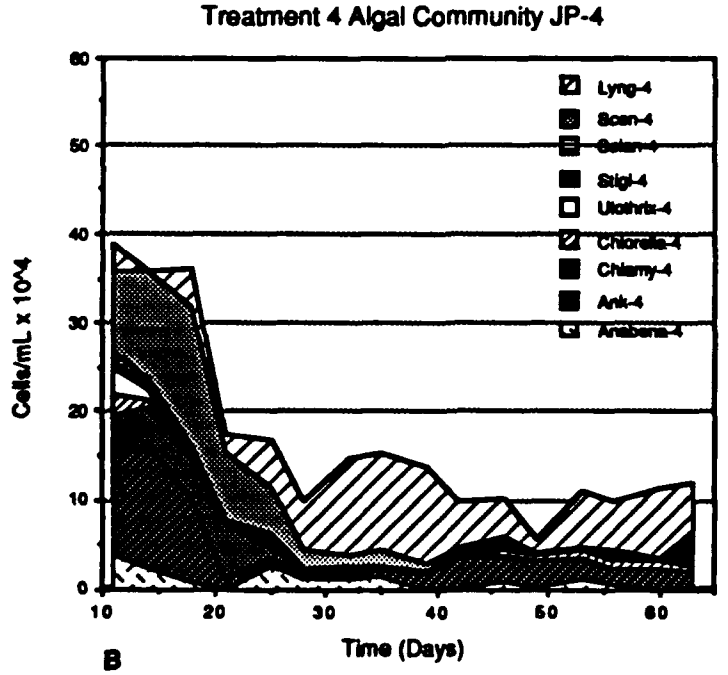
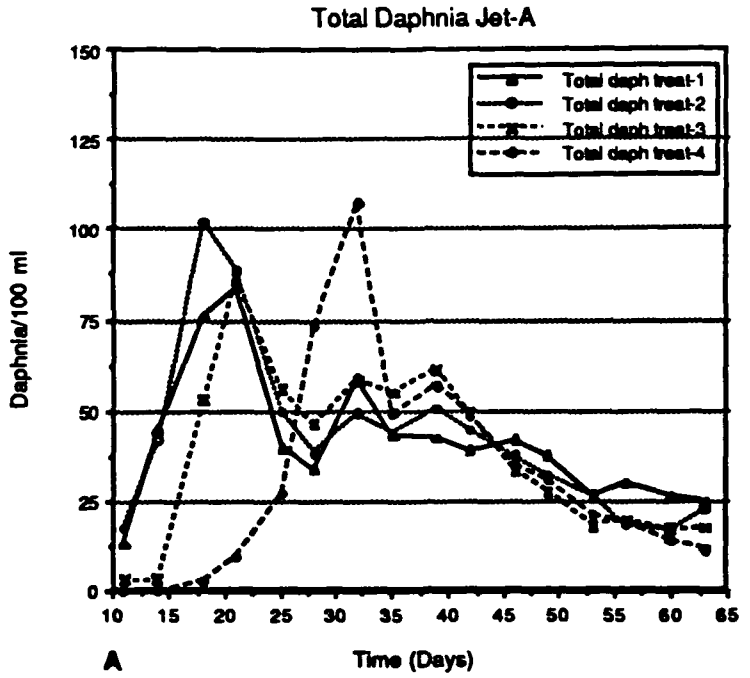


FIG. 1--Comparison of algal population dynamics-highest treatment.



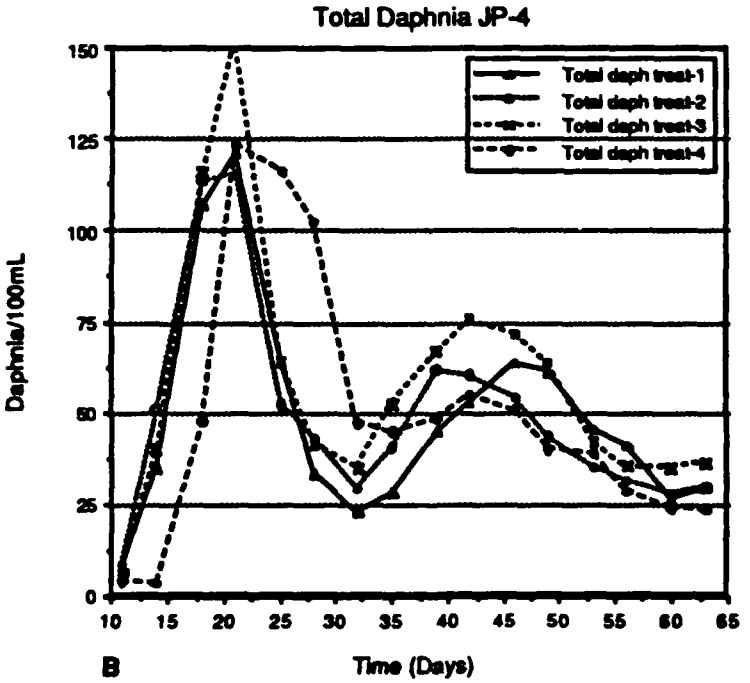


FIG. 2--Daphnid population dynamics.

The experiment using JP-4 as a toxicant (B) did not see the increase in ostracods until between days 50-55, approximately ten days later. Consequently, the total numbers of ostracods observed were not as high in the JP-4 microcosms. Note that the order of densities in the Jet-A experiment followed a dose response pattern, as did the JP-4 experiment, even with the lower total numbers. Conventional analysis did not demonstrate significance, however non-metric clustering did indicate the importance of the ostracods in determining clusters in both sets of microcosm experiments.

Philodina Population Dynamics. Philodina did not become prevalent in the microcosms until the second half of the experiment. One of the major problems was the inherent variability in the sampling and in the replicates. Organisms that reproduce rapidly can show large differences in population sizes during the course of a sampling day. Although, in the later stages of the microcosm experiments the dosed systems had a generally larger number of the rotifers, the results were not statistically significant using conventional IND plots. However, using cluster analysis, Philodina were also determined to be an important variable in defining clusters. This held true for both the Jet-A and JP-4 experiments.

Comparisons of pH dynamics of the Jet-A and JP-4 Experiments. Unlike the biotic variables, pH did reflect some of the the oscillations detected by the cluster analysis (Fig. 4). In both the Jet-A and the JP-4 experiments the highest concentrations demonstrated a statistically significant difference, determined by the interval of non-significant difference during the first 30 days of the experiment. The second oscillation, between days 45 and 50, is not as clear since only one sampling date demonstrated the statistically significant difference. Type II error becomes a concern with so many comparisons, even with the corrections incorporated into the IND plots.

Photosynthesis/Respiration Ratio. The photosynthesis/respiration ratio reflects the oscillations seen in pH and the clustering analysis for the first 30 days and then only for the Jet-A water soluble fraction. In the Jet-A experiment, a second deviation from the IND plot was noted in the period corresponding to the second oscillation, but the result is difficult to distinguish from a type II error. In the JP-4 experiment, the IND plots are large, reflecting the variance in those sampling days. As an "emergent property", it is not clear if the P/R ratio provides any more information in this experiment than the clustering based upon the biotic components.

Oscillations in Community Dynamics Observed in both the Jet-A and the JP-4 Experiments. The Jet-A and the JP-4 SAM experiments both displayed a series of oscillations; revealed by the three clustering techniques employed in the analysis (Fig. 5). The first oscillation, as defined by Cosine Distance common to each experiment, is due to the

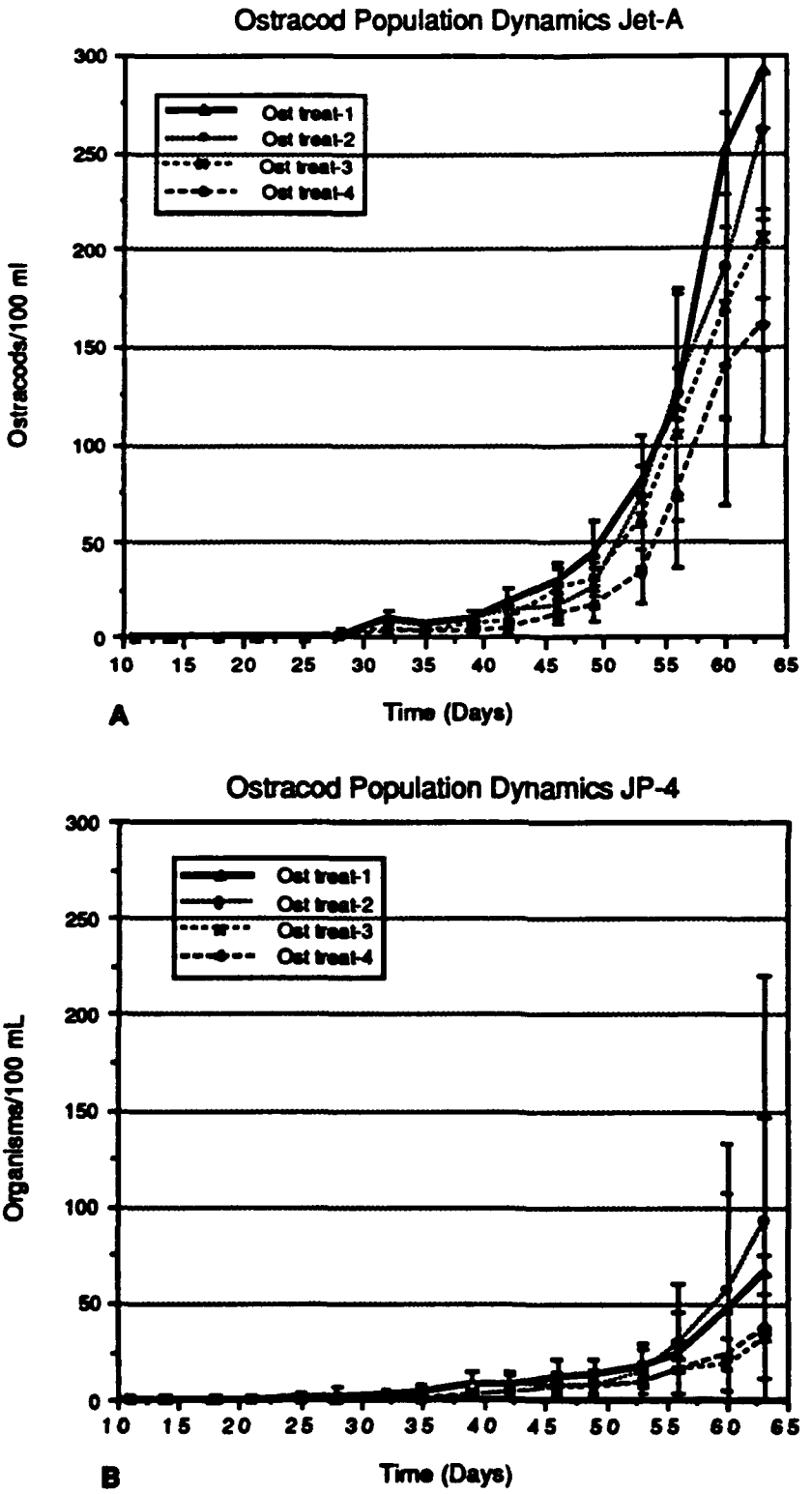
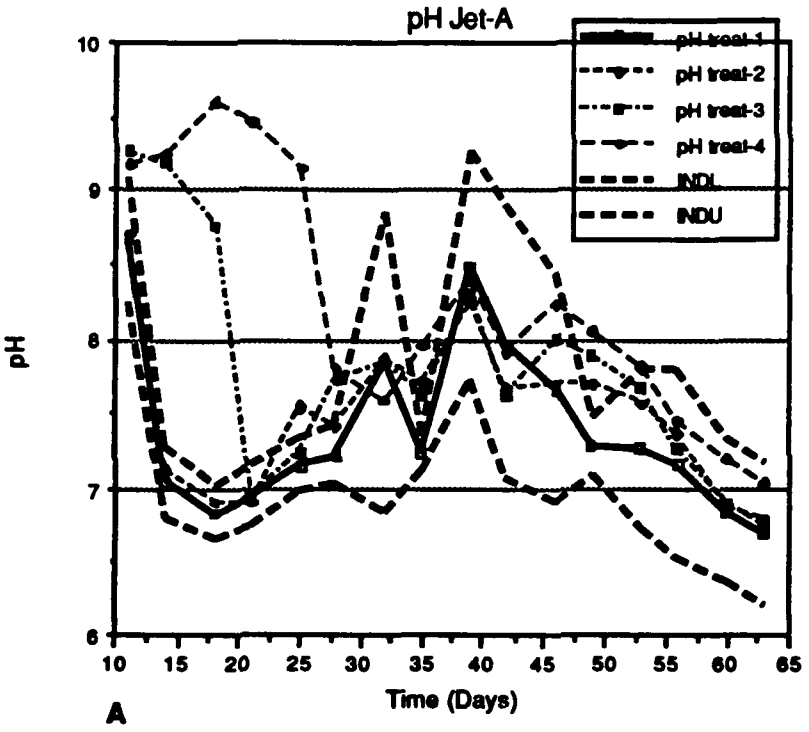


FIG. 3--Ostracod population dynamics.



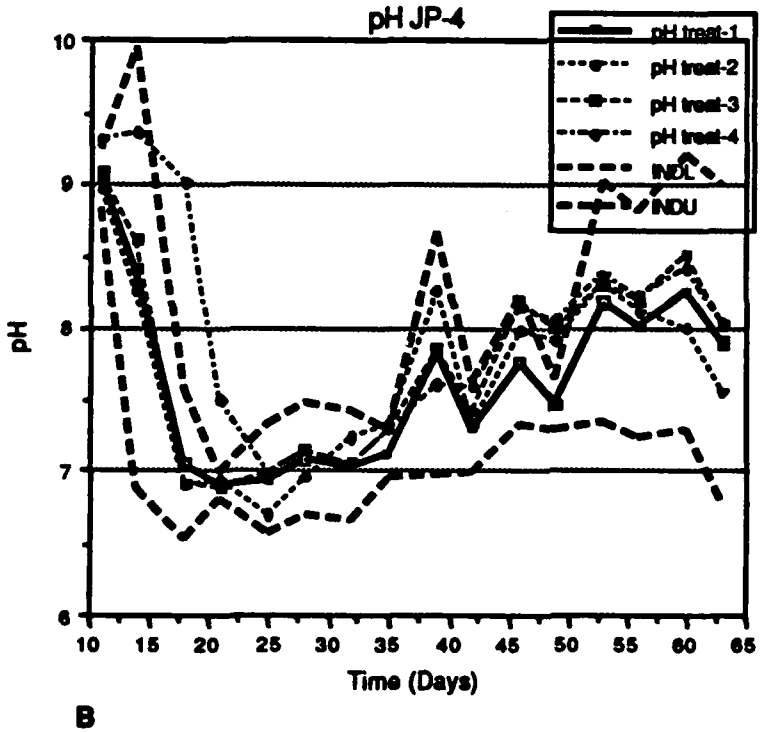


FIG. 4--Comparisons of pH during the SAM studies.

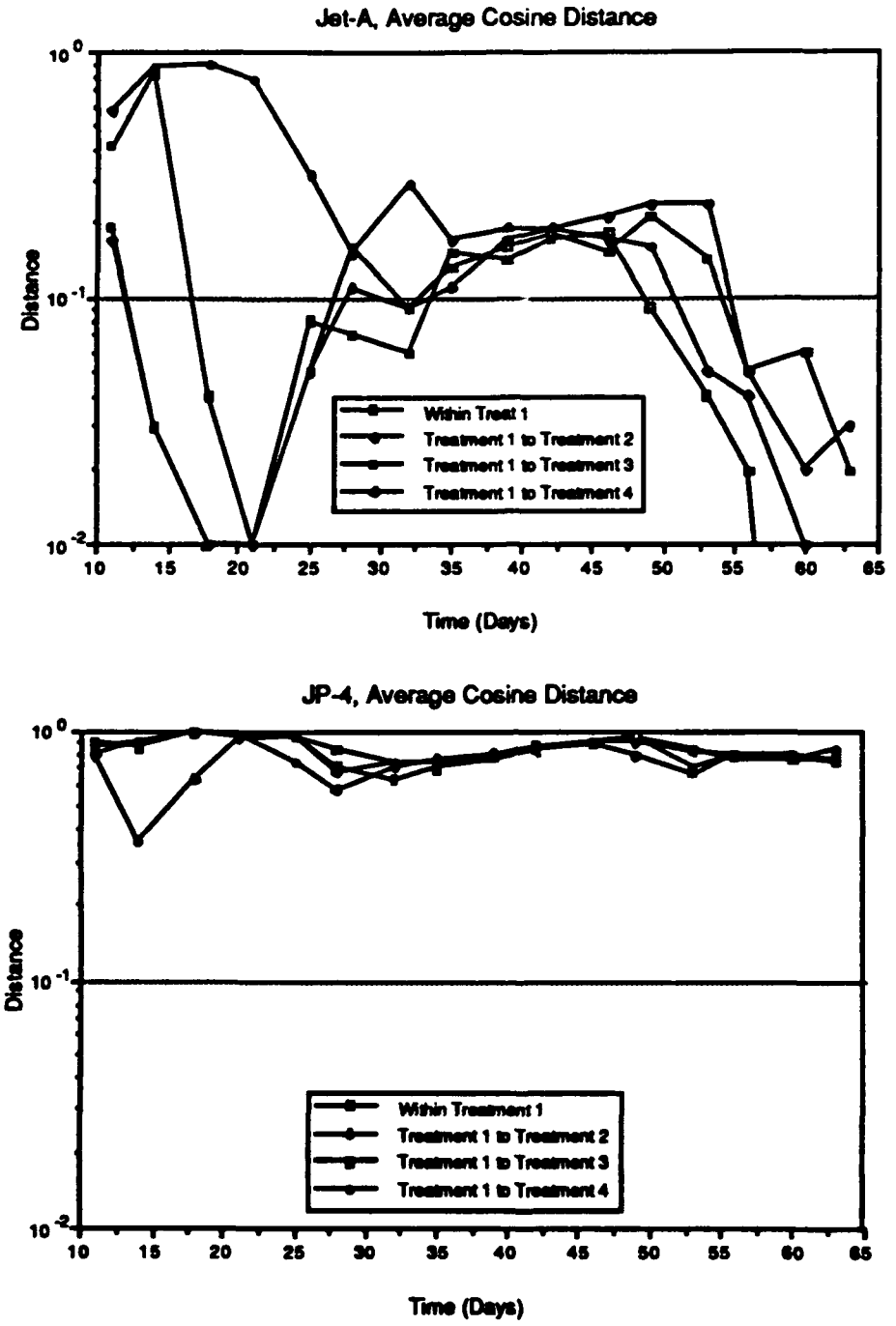


FIG. 5--Cosine Distances of the Treatment 1 to the dosed treatments in the Jet-A and the JP-4 SAMs.

interaction of the daphnid population and the algae. The result is statistically significant, as determined by the goodness-of-fit confidence level, graphed by day in Fig. 6. In both experiments, the oscillation is within the first 30 days of the SAM time-line. Interestingly, the magnitude of the first oscillation, as determined by Cosine Distance, is less in the JP-4 experiment, possibly reflecting the reduced acute and chronic toxicity of the mixture.

A second series of oscillations, as measured by Cosine Distance, occur in the last thirty days of each experiment. Again the oscillations are statistically significant.

TABLE 2. Variable ranking by success in determining clusters as defined by nonmetric clustering. Variables such as Ankistrodesmus and the Daphnia classes ranked highly in the course of this study. However, reliance on any particular organism or a small combination of variables would inadequately describe the dynamics of the system.

Jet-A		JP-4	
Variable	Ranked	Variable	Ranked
Ankistrodesmus	12	Chlorella	8
M. Daphnia	11	S. Daphnia	8
Chlorella	9	Ankistrodesmus	6
Scenedesmus	7	Scenedesmus	5
S. Daphnia	6	Philodina	5
L. Daphnia	5	M. Daphnia	4
Ostracod	4	Lyngbya	4
Philodina	4	L. Daphnia	3
Selenastrum	4	Ostracod	3
Lyngbya	3	Selenastrum	3
Ulothrix	1		

The participants in the community that contribute to these oscillations are slightly different judging by the table of important variables (Table 2). Unfortunately, the length of the SAM protocol is not sufficient to conduct an analysis of the period and amplitude of the oscillations. Another complication in examining the results is the difficulty in making direct comparisons between experiments. Although the Cosine Distance may be the same, the orientation of the angle can be quite different.

DISCUSSION

First, the apparent recovery or movement of the dosed systems towards the reference or treatment 1 case may be an artifact of our measurement systems that allow the n-dimensional data to be represented in a two dimensional system. In an n-dimensional sense, the systems may be moving in opposite directions and simply pass by similar coordinates during certain time intervals. Positions may be similar but the n-dimensional vectors describing the movements of the systems can be very different. A representation of these dynamics is presented in Fig. 7. The two systems intersect, although the vectors are quite different.

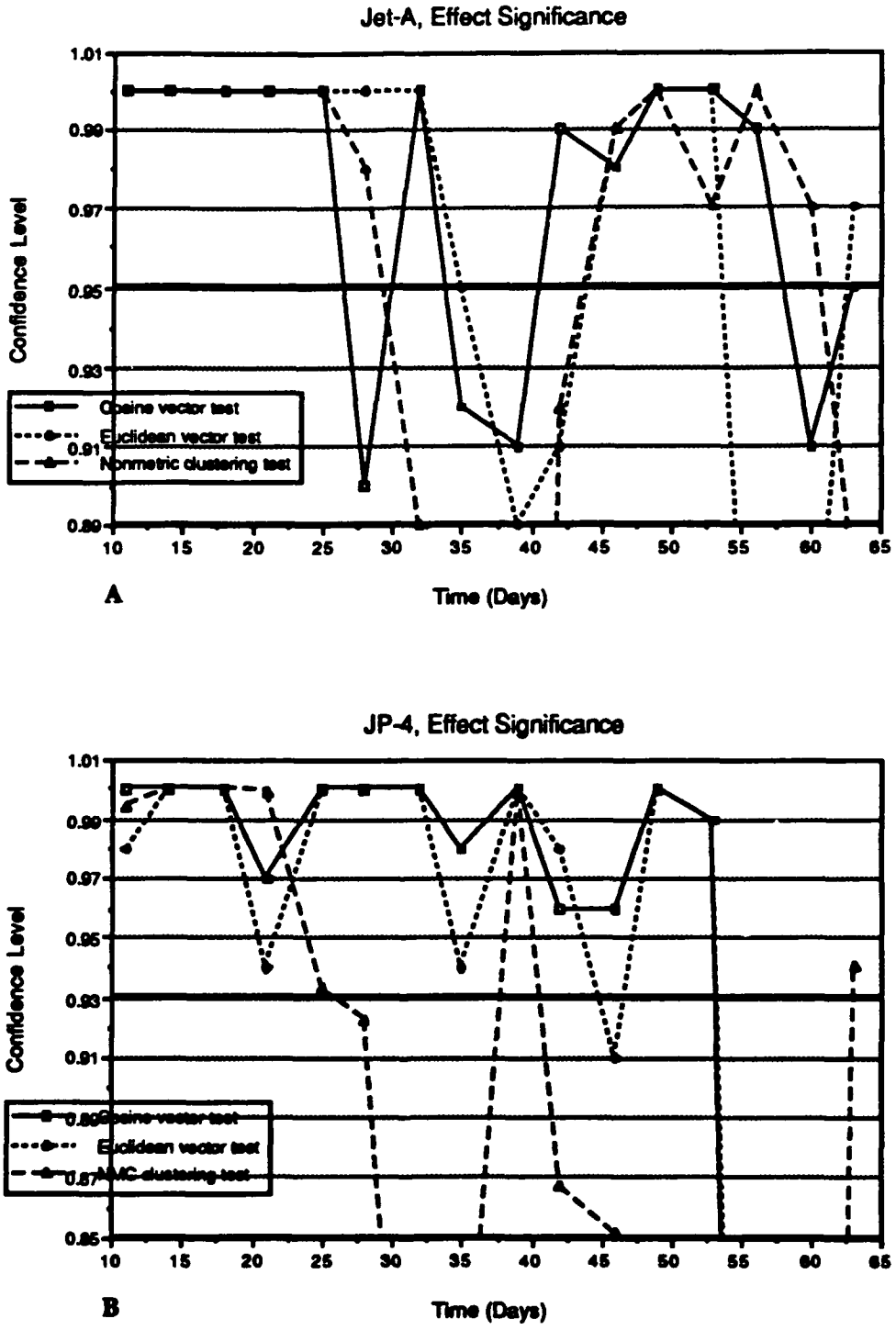


FIG. 6--Significance of the association analysis of the 4 Treatments in the Jet-A and the JP-4 SAMs.

The apparent recoveries and divergences may also be artifacts of our attempt to choose the best means of collapsing and representing n-dimensional data into a two or three dimensional representation. In order to represent such data it is necessary to project n-dimensional data into three or less dimensions. As information is lost as the shadow from a cube is projected upon a two dimensional screen, a similar loss of information can occur in our attempt to represent n-dimensional data. Not every divergence from the reference treatment may have a cause directly related to it in time. Differentiating those events from those due to degradation products or other perturbations is challenging.

Not only may system recovery be an illusion, but there are strong theoretical reasons that seem to indicate that recovery to a reference system may be impossible or at least unlikely. In fact, systems that differ only marginally in their initial conditions and at levels probably impossible to measure are likely to diverge in unpredictable manners. May and Oster (1978) in a particularly seminal paper investigated the likelihood that many of the dynamics seen in ecosystems that are generally attributed as chance or stochastic events are in fact

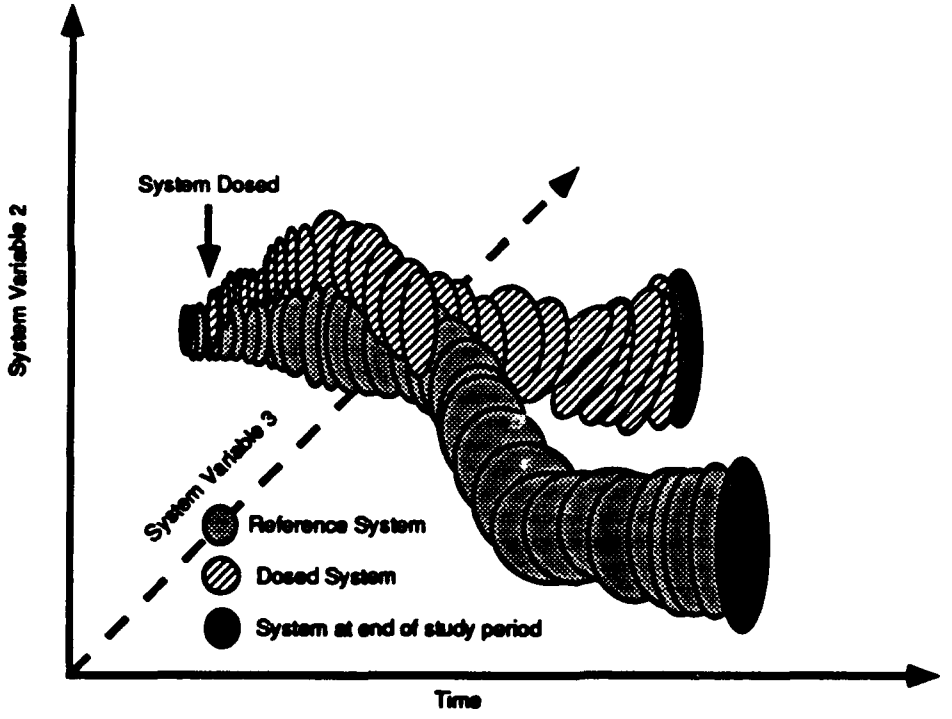


FIG. 7--Visualization of ecosystem dynamics to reflect a possible interpretation of the impacts of the jet fuels.

deterministic. In fact, simple deterministic models of populations can give rise to complex dynamics. Using equations resembling those used in population biology, bifurcations occur resulting in several distinct outcomes. Eventually, given the proper parameters, the system appears chaotic in nature although the underlying mechanisms are completely deterministic. Obviously, biological systems have limits, extinction being perhaps the most obvious and best recorded. Another ramification is that the noise in ecosystems and in sampling may not be the result of

a stochastic process but the result of underlying deterministic, but chaotic relationships.

These principals also apply to spatial distributions of populations as recently reported by Hassell et al. (1991). In a study using host-parasite interactions, a variety of spatial patterns were developed using the Nicholson-Bailey model. Host-parasite interactions demonstrated dynamics ranging from static 'crystal lattice' patterns, spiral waves, chaotic variation, or extinction with the appropriate alteration of only three parameters within the same set of equations. The deterministically determined patterns could be extremely complex and not distinguishable from stochastic environmental changes.

Given the perhaps chaotic nature of populations it may not be possible to predict species presence, population interactions, or structural and functional attributes. Kratz et al. (1987) examined the spatial and temporal variability in zooplankton data from a series of five lakes in North America. Much of the analysis was based on limnological data collected by Brige and Juday from 1925 to 1942. Copepods and cladocera, except *Bosmina*, exhibited larger variability between lakes than between years in the same lake. Some taxa showed consistent patterns among the study lakes. They concluded that the controlling factors for these taxa operated uniformly in each of the study sites. However, in regards to the depth of maximal abundance for calanoid copepods and Bosmina, the data obtained from one lake had little predictive power for application to other lakes. Part of this uncertainty was attributed to the intrinsic rate of increase of the invertebrates with the variability increasing with a corresponding increase in r_{max}. A high r_{max} should enable the populations to accurately track changes in the environment. Katz et al suggest that these taxa be used to track changes in the environment. Unfortunately, in the context of environmental toxicology, the inability to use one "reference" lake to predict the non-dosed population dynamics of these organisms in another eliminates comparisons of the two systems as measures of anthropogenic impacts.

A better strategy may be to let the data and a clustering protocol identify the important parameters in determining the dynamics of and impacts to ecological systems. This approach has been recently suggested independently by Dickson et al. (1992) and Matthews and Matthews (Matthews et al. 1991b, Matthews and Matthews 1991). This approach is in direct contrast to the more usual means of assessing anthropogenic impacts. One classical approach is to use the presence or absence of so called indicator species. This assumes that the tolerance to a variety of toxicants is known and that chaotic or stochastic influences are minimized. A second approach is to use hypothesis testing to differentiate metrics from the systems in question. This second approach assumes that the investigators know *a priori* the important parameters to measure. Given that in our relatively simple SAM systems that the important parameters in differentiating non-dosed from dosed systems change from sampling period to sampling period, this assumption can not be made. Classification approaches such as nonmetric clustering or the canonical correlation methodology developed by Dickson et al, eliminates these assumptions.

These results presented in this report and by others reviewed above and the implications of chaotic dynamics suggest that reliance upon any one variable or an index of variables may be an operational convenience that may provide a misleading representation of pollutant effects and associated risks. The use of indices such as diversity and the Index of Biological Integrity have the effect of collapsing the dimensions of the descriptive hypervolume. Indices, since they are

composited variables, are not true endpoints. The collapse of the dimensions that are composited tends to eliminate crucial information, such as the variability in the importance of variables. The mere presence or absence and the frequency of these events can be analyzed using techniques such as nonmetric clustering that preserve the nature of the dataset. A useful function was certainly served by the application of indices, but the new methods of data compilation, analysis and representation derived from the Artificial Intelligence tradition can now replace these approaches and illuminate the underlying structure and dynamic nature of ecological systems.

The implications are important. Currently, only small sections of ecosystems are monitored or a heavy reliance is placed upon so called indicator species. These data suggest that to do so is dangerous, may produce misleading interpretations resulting in costly error in management and regulatory judgments. Much larger toxicological test systems are currently analyzed using conventional statistical methods on the limit of acceptable statistical power. Interpretation of the results has proven to be difficult, if not confusing. Application of the approach and tools that proved successful in revealing the complex dynamics of these small microcosms should prove useful in analyzing larger toxicological test systems and field research.

CONCLUSIONS

(1) In both of the experiments, multiple oscillations of the dosed treatment groups away from the reference treatment were observed using multivariate statistics. The first oscillation is due to the differential impact of the WSF of the jet fuels to the algae-daphnid population dynamics. The following oscillations, although statistically significant and seen in both experiments, is not as clear cut.

The divergence of the second oscillation may be due to two separate mechanisms.

(a) A fluctuation due to the initial stress has occurred, but in such a fashion that an incompletely dampened oscillation repeats. There has been no fundamental alteration in the functioning of the ecosystem, and the oscillations are a result of the inherent time lags and stochastic factors governing the dynamics of the system.

(b) A fundamental aspect of the ecosystem has been altered so that the repeated oscillations reflect the persistence of the impact. An alteration in the detritus quality or in the community involved in the recycling of detritus may have long term impacts as other nutrients become limiting in the system. Nutrients are at low levels during the second 30 days of a typical SAM experiment. This possibility could include a fundamental and long lasting effect upon the system, contrary to the first mechanism.

(2) A combination of multivariate analyses appear to be useful and illuminating in assessing the long term dynamics of these systems. Each has strengths that make multivariate analysis a strong methodology with powerful advantages to conventional univariate methods.

(3) Although simple systems, the SAM experiments exhibits complex dynamics and behaviors. The protocol results in a persistent system with good replicability within an experiment, even with complex species interactions.

(4) Techniques that allow the reduction and visualization of even these relatively simple multispecies toxicity tests should contribute to our understanding of system dynamics and improve hazard assessment.

ACKNOWLEDGEMENTS

This research was funded by United States Air Force Office of Scientific Grant No. AFOSR-91-0291 DEF.

REFERENCES

ASTM D2887. Standard Test Method for Boiling Range Distribution of Petroleum Fractions by Gas Chromatography, 1988 Annual book of ASTM Standards, Vol. 5.02, pp 506-513. American Society for Testing and Materials, Philadelphia.

ASTM D3710. Standard Test Method for Boiling Range Distribution of Gasoline and Gasoline Fractions by Gas Chromatography, 1988 Annual Book of ASTM Standards, Vol. 5.03, pp 78-88. American Society for Testing and Materials, Philadelphia.

ASTM E 1366-91. 1991. Standard Practice for the standardized aquatic microcosm: fresh water, Vol 11.04. pp 1017-1051. American Society for Testing and Materials, Philadelphia.

Conquest, L.L. and Taub, F.B. 1989. Repeatability and reproducibility of the Standard Aquatic Microcosm: Statistical properties. In U.M. Cowgill and L.R. Williams, eds., *Aquatic Toxicology and Hazard Assessment: 12th Volume, ASTM STP 1027*. American Society for Testing and Materials, Philadelphia, PA, pp. 159-177.

Crow, M.E. and Taub, F.B. 1979. Designing a microcosm bioassay to detect ecosystem level effects. *Intern. J. Environmental Studies*. 141-147.

Dickson, K.L., Waller, W.T., Kennedy, J.H. and Ammann, L.P. 1992. Assessing the relationship between ambient toxicity and instream biological response. *Env. Tox. Chem.* 11:1307-1322.

Hassell, M.P.H, Comins, N. and May, R.M. 1991. Spatial structure and chaos in insect population dynamics. *Nature* 353:255-258.

Johnson, A.R. 1988a. Evaluating ecosystem response to toxicant stress: a state space approach. *Aquatic Toxicology and Hazard Assessment: 10th Volume, ASTM STP* 971, W.J. Adams, G.A. Chapman and W.G. Landis Eds., American Society for Testing and Materials, Philadelphia, pp. 275-285.

Johnson, A.R. 1988b. Diagnostic variables as predictors of ecological risk. *Environmental Management* 12:515-523.

Katz, T.K., Frost, T.M. and Magnuson, J.J. 1987. Inferences from spatial and temporal variability in ecosystems: Long-term zooplankton data from lakes. *Amer. Nat.* 129:830-846.

Kersting, K. 1984. Development and use of an aquatic micro-ecosystem as a test system for toxic substances. Properties of an aquatic micro-ecosystem IV. *Int. Revue ges. Hydrobiol.* 69:567-607.

Kersting, K. 1985. Properties of an aquatic micro-ecosystem V. Ten years of observations of the prototype. *Verh. Internat. Verein. Limnol.* 22:3040-3045.

Kersting, K. 1988. Normalized ecosystem strain in micro-ecosystems using different sets of state variables. *Verh. Internat. Verein. Limnol.* 23:1641-1646.

Kindig, A.C., Loveday, L.C. and Taub, F.B. 1983. Differential sensitivity of new versus mature synthetic microcosms to streptomycin sulfate treatment. *Aquatic Toxicology and Hazard Assessment: Sixth Symposium, ASTM 802.* W.E. Bishop, R.D. Cardwell and B.B. Heidolph Eds. American Society for Testing and Materials, Philadelphia, pp. 192-203.

Landis, W.G., Haley, M.V. and Chester, N.A. 1993. The use of the standardized aquatic microcosm in the evaluation of degradative bacteria in reducing impacts to aquatic ecosystems. In W.G. Landis, J. Hughes and M. Lewis., eds., *In press, Environmental Toxicology and Risk Assessment: First Volume, ASTM STP -1179.* American Society for Testing and Materials, Philadelphia, PA.

Landis, W.G., Matthews, R.A., Markiewicz, A.J., Shough, N.J. and Matthews, G.B. 1993. Multivariate analysis of the impacts of turbine fuel using a standard aquatic microcosm toxicity test. *J. Env. Sci.* In Press.

Matthews, G.B. and Hearne, J. 1991. Clustering without a metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 13:175-184.

Matthews, G.B. and Matthews, R.A. 1991. A model for describing community change. In *Pesticides in Natural Systems: How Can Their Effects Be Monitored? Proceeding of the Conference,* Environmental Research Laboratory/ORD, Corvallis, OR, EPA 9109/9-91/011.

Matthews, G.B., Matthews, R.A. and Hachmoller, B. 1991b. Mathematical analysis of temporal and spatial trends in the benthic macroinvertebrate communities of a small stream. *Canadian Journal of Fisheries and Aquatic Sciences.* 48:2184-2190.

Matthews, R.A., Matthews, G.B. and Ehinger, W.J. 1991a. Classification and ordination of limnological data: a comparison of analytical tools. *Ecological Modeling.* 53:167-187.

May, R.M. and Oster, G.F. 1978. Bifurcations and dynamical complexity in simple ecological models. *Amer. Nat.* 110:573-599.

Taub, F.B. 1969. Gnotobiotic models of freshwater communities.*Verh Internat. Verein. Limnol.* 17, 485-496.

Taub, F.B. 1976. Demonstration of pollution effects in aquatic microcosms. *Intern J. Environmental Studies.* 10, 23-33.

Taub, F.B. 1988. Standardized aquatic microcosm - development and testing. *Aquatic Ecotoxicology* II.

Taub, F.B. 1989. Standardized aquatic microcosms. *Environm. Sci. Technol.* 23:1064-1066.

Taub, F.B., Kindig, A.C. and Conquest, L.L. 1987. Interlaboratory testing of a standardized aquatic microcosm. In W.J. Adams, G.A. Chapman and W.G. Landis, eds., *Aquatic Toxicology and Hazard Assessment: 10th Volume, ASTM STP 971.* American Society for Testing and Materials, Philadelphia, PA, pp. 385-405.

Taub, F.B., Kindig, A.C., Conquest, L.L. and Meador, J.P. 1988. Results of the interlaboratory testing of the Standardized Aquatic Microcosm protocol. In G. Suter and M. Lewis, eds., *Aquatic Toxicology and Hazard Assessment: Eleventh Symposium, ASTM.* American Society for Testing and Materials, Philadelphia, PA.

Taub, F.B. and Crow, M.E. 1978. Loss of a critical species in a model (laboratory) ecosystem. *Verh. Internat. Verein. Limnol.* 1270-1276.

Taub, F.B. , Crow, M.E. and Hartmann, H.J. 1980. Responses of aquatic microcosms to acute mortality. *Microcosms in Ecological Research.* Giesy, J.P. Jr. Technical Information Center, U. S. Department of Energy. Washington, D.C., 513-535.

Westendorf, R.G. 1986. Performance aspects of volatile organics analysis by purge and trap capillary column gas chromatography with flame ionization detectors. Tekmar Technical Papers, Tekmar Company, Cincinnati, Ohio.

Nonmetric Clustering and Association Analysis: Implications for the Evaluation of Multispecies Toxicity Tests and Field Monitoring*

G. B. Matthews
W. G. Landis
R. A. Matthews

April 29, 1993

Contents

1 Introduction **2**

2 Machine Learning **4**

3 Nonmetric Clustering **5**
3.1 Numerically ranking conceptual descriptions 6
3.2 Integrating qualitative and quantitative data 8

4 Association Analysis: a Significance Test from the Clustering **10**

5 Implications for Ecological and Ecotoxicological Tests **10**

6 Future Work: Dynamic Ecosystem Change **12**

7 Conclusion **13**

Abstract

Many techniques developed by computer scientists in the field of artificial intelligence (AI) are currently being used as standard, state-of-the-art technology. These techniques have proven their value and validity in medicine, geology, agronomy, and astronomy time and again, often beating human experts at their own game. We present here an analysis tool

*This research is supported by USAFOSR Grant No. AFOSR-91-0291 DEF.

for multispecies data based on nonmetric clustering, an AI technique developed specifically to aid in the interpretation of complex ecological data sets. This technique uses AI search to find an appropriate and meaningful characterization of a multivariate system. After appropriately characterizing the system in this fashion, the relationship between this characterization of the system and the critical environmental variables (pollution, toxicity, *etc.*) can be quantitatively analyzed to aid in the assessment of the effects of the environment on the system. *A priori* endpoints or indices are not necessary; the data are allowed to determine the variables that best separate treatment from controls. We have now tested this methodology over a series of multispecies toxicity tests using a variety of stressors. During the initial blind testing the methodologies could pick treatment groups with high accuracy. When knowledge of treatment group is available, oscillations in the similarity of the treatments to the controls are apparent.

Much recent debate in toxicological studies has focussed on appropriate endpoitns for multispecies toxicity tests and biomonitoring schemes. We suggest that the search for endpoints appropriate to the entire field of toxicity testing is a fruitless search. We recommend instead an approach that standardizes the common sense approach: different situations, even within a single experiment, call for different endpoints. Typically, the toxicologist, if called upon for an expert opinion, will examine multivariate data, and extract from that data a few critical species. The behavior of these species will give an adequate (though perhaps not complete) picture of the toxic effects. Which species are selected, and whether it is their mortality, behaviro, or biomass that is important, will always vary from case to case. We call, therefore, for more research into the automation of the process typically performed by the expert. The selection of species, as well as other parameters, as significant for a particular experiment or field study, can be done automatically by computer algorithms. To be blind to the utility of these tools in the field of toxicology is to work by hand, over and over again, problems which could be solved in a twinkling with their aid.

1 Introduction

It has become a shibboleth in modern ecotoxicology that the field cannot progress until ecologically significant endpoints are defined. Something along the lines of an ecosystem level functional index, it is presumed, would be ideal, telling us what numbers to measure, which mathematical formulae to use to boil them down, and where the cutoff point is between healthy systems and troubled ones. This would introduce "objectivity" into what is now done with an intuitive assessment by a human expert. The reality, however, is that the state of an ecological community cannot possibly be captured on any linear scale, on the one hand, and, on the other, that an approach to assessment using the traditional human "best judgement" is doomed to failure by the innately incomprehensible

complexity of the analysand. Fortunately, there is a middle ground for dealing with complex systems. In other scientific domains, practitioners have long realized their impotence in the face of massive multivariate data, and have resorted to automated computerized tools for image processing, pattern recognition, and dimensionality reduction. These tools are in widespread use, for example, in medicine, astrophysics, particle physics, meteorology, and geology. The key to their success is that the human expert and the software tool are partners in the exploration of the data. The computer by itself, of course, has no semantic understanding of the data. But, equally, the unaided human is blind to the patterns implicit in the data. Increasingly sophisticated data visualization and analysis tools are available on today's powerful desktop workstations, and the practitioner who does not use them will soon be left behind.

Much of the work in computer-aided data exploration, however, has the wrong focus for ecotoxicology. Data sets generated, for example, by meteorological models of a thunderstorm, typically have millions of data points densely scattered through a well-defined three-dimensional model. The complexity is in the sheer number of data points and their interactions. In ecologically interesting situations, on the other hand, only a few dozen or hundred data points are in hand, from widely separated places in space and time, and each point records data on dozens or hundreds of species. This results in a relatively small number of points scattered through the huge volume of n-dimensional space (where n is the number of different species counted). Even a modest number of dimensions raises severe problems for conventional analysis techniques, and human intuition. For example, if some large number of points is scattered uniformly over a 10-dimensional hypersphere with radius one, then a hypersphere inside, of radius 3/4, will contain only 5% of the points. Clearly, sampling 10 or higher dimensional space can miss important things. Further, a lot of the time data points are missing, or incomplete.

The nature of the problem is that usually we have *too much* information. Ten or twenty sampling points with, perhaps, fifty species, is underdetermined. There is no way to draw meaningful conclusions about the nature of the community as a whole (all fifty dimensions), from the smattering of points. What is required is *data reduction*, the dimensionality of the data has to be brought down to the point where ten or twenty points can tell us something. One methodology for this is based on *projections* of the data, such as factor analysis, principal components analysis, correspondence analysis, or, more generally, projection pursuit (Huber, 1985). There are many algorithms for finding good projections, and even a suggestion that *all* projections be examined in a "grand tour" of the data (Asimov, 1985). However, rotating at about 10° per second, a reasonable speed for careful observation, a grand tour of only four dimensions would take about three hours (Huber, 1985), and so computer-aided projections are the only real alternative.

While such projections are valuable in reducing the dimensionality of the data, they all suffer from a problem of comprehensibility. Since arbitrary linear

and nonlinear transformations of the data matrix are allowed, the meaning of the resulting two-dimensional projection can be obscure, and difficult for human intuition to fathom.

The tradition of machine learning (ML), within artificial intelligence, has been addressing these problems for some time. The goal of an ML system is, not only to identify patterns in the data, but to come up with an efficient and intuitive characterization of them. Efficient and intuitive, in this context, imply that the characterization is not unnecessarily complex, that it uses simple logical combinations of descriptions rather than mathematical formulae, and that it is expressed in terms of attributes that are not contrived. This has been formulated as the *comprehensibility postulate*:

> The results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single "chunks" of information, directly interpretable in natural language, and should relate quantitiative and qualitative concepts in an integrated fashion (Michalski, 1983).

It is the primary failing of traditional statistical approaches, as well as the "neural net" approach, to solving ML problems that they ignore the comprehensibility postulate. In this paper, we present nonmetric clustering, a specialization of ML, faithful to the comprehensibility postulate, which we have been employing fruitfully on a wide variety of ecosystems. After its details are explained, some consequences for environmental policy making are outlined.

2 Machine Learning

As a simple example, consider the data in Table 1 (Quinlan, 1983). In this set, we are given three "positive" individuals and five "negative" individuals and their characteristics on three attributes. The problem is to come up with a means of distinguishing the "positives" from the "negatives" based on height, hair, and eye color. There are many possible ways of distinguishing them, but one nice one might be:

> Positives either have red hair, or blond hair and blue eyes.
> Negatives either have dark hair, or blond hair and brown eyes.

There are several things to notice about this characterization of the positives and negatives. First, the data are both categorical and numeric. The beauty of ML approaches to these problems is that they apply equally well to either kind of data. To make a regression, or linear discriminant, categorical data would have to be numerically coded somehow. In an ML approach, numeric attributes, such as height, are simply recoded into a number of discrete bins,

Height	Hair	Eyes	Class
short	blond	blue	+
tall	blond	brown	−
tall	red	blue	+
short	dark	blue	−
tall	dark	blue	−
tall	blond	blue	+
tall	dark	brown	−
short	blond	brown	−

Table 1: Data set problem for identifica and characterization.

such as small, medium, and large. Such categories can be as fine or as coarse as desired, and in all events are more comprehensible than an uninterpreted number. Second, not all the original attributes are used in the description. Height, it turns out, is superfluous, and is omitted from the description. Third, compound descriptions are created using logical operations, "and" "or" and "not", rather than mathematical formulae. A linear discriminant, for example, describes by adding up numbers and then determining if the result is greater or smaller than some cutoff point. The logical descriptions are much more natural and intuitive for humans, and lead to understanding of the data in a way that mathematical combinations cannot. Fourth, even with only three attributes and eight points, there are a lot of different logical descriptions that have to be considered to get the best, or even a good, one. With real data sets the combinatorial complexity of finding a description would rapidly swamp a human investigator. A computer aid is essential. Fifth, no artificial attributes are used. The use of "indices" or "ordination" techniques attempts to introduce a new attribute, defined mathematically in terms of the original ones, and then use the values of these indices or components to describe the classes. The ML description uses the same attributes (height, hair, and eyes) that were used in the design of the sampling program, and thus, the description of the classes will have direct meaning to the investigator, without the need to learn a new vocabulary. Such descriptions, which use simple logical combinations of the original attributes, are called "conceptual" descriptions (Michalski and Stepp, 1983).

3 Nonmetric Clustering

Nonmetric clustering (NMC) is an ML tool designed to search for conceptual descriptions of ecological data sets. The NMC methodology has been implemented in a computer program called Riffle (Matthews and Hearne, 1991). Unlike the simple example above, Riffle does not work from a preexisting set of class labels (such as + and −). Given a data set, Riffle attempts to two things simultane-

C1	A	B	C	D	E	F	C2
+	1	2	1	2	2	1	+
+	1	2	2	1	1	2	−
+	1	2	1	2	2	1	+
−	2	1	2	1	1	2	−
−	2	1	1	2	2	1	+
−	2	1	2	1	1	2	−

Table 2: Synthetic data for nonmetric clustering, with two possible clusterings.

ously: Group the points into clusters (classes), and find the simplest possible conceptual description of those clusters. Since the points are not previously assigned to classes, Riffle is free to give the points any class label at all. However, the class labels must be such that they can be simply captured in a conceptual description, based on the original attributes (measured parameters), and, further, such that they, in turn, capture as much information as possible about the original attributes.

Consider the synthetic data in Table 2, where six points have been sampled for six attributes. One potential clustering, denoted C1, has two simple conceptual descriptions, each based on a single attribute, either A or B. C, D, E and F can be regarded as superfluous for this clustering. Another potential clustering, denoted C2, also has simple characterizations, but in terms of attributes C, D, E, and F, with A and B as superfluous. While both clusterings have simple conceptual descriptions, C2 should be preferred because it captures more information about the points than C1. One way to express this algorithmically is that there are *more* good conceptual descriptions of the classes in C2 than there are of the classes in C1. The computer program Riffle will prefer C2 to C1 for this reason.

To find the best clustering possible, for a given data set, the algorithm works by examining a great number of possible clusterings, like C1 and C2, above, and numerically ranks their conceptual adequacy. All data points are repeatedly reassigned to clusters, and then the conceptual association between clusters and attributes is reevaluated. When an assignment of points to clusters is found that outranks all others, it is reported as the most natural clustering.

We will now briefly discuss how conceptual adequacy is ranked, and also make some remarks on the particular strategy used in Riffle to convert numeric to categorical variables.

3.1 Numerically ranking conceptual descriptions

To begin with, assume all attributes are categorical. Nonmetric clustering measures the association between a clustering (which, itself, is a categorical variable) and another categorical variable by means of a contingency table test. A fre-

	A1	A2	A3
B1	5	3	1
B2	1	4	2
B3	7	0	5

Table 3: A contingency table to illustrate calculation of Guttman's λ.

quency table of cluster-number *vs.* categorical-value is set up, and the number of data points in each cell is counted in order to measure the association between cluster and variable. The most famous contingency table test is probably the χ^2 test, but the χ^2 test has some undesirable properties when it comes to interpretation and comprehensibility. Nonmetric clustering uses Guttman's λ to measure the association in the table (Goodman and Kruskal, 1954; Goodman and Kruskal, 1959; Goodman and Kruskal, 1963; Goodman and Kruskal, 1972).

Guttman's λ is a measure defined on the basis of "optimal predictions". Consider, for instance, the contingency table represented in Table 3. Twenty-eight individuals have been sampled, and their values on attributes A and B have been tabulated. For concreteness, A can be regarded as "height" and B as cluster-number. A larger sample size would always be desirable, but we have no recourse other than to regard the proportion of points found in any cell as the best estimate of the probability of finding a new point also to be in that cell. Now suppose we need to predict which value on attribute B a new sample is likely to have. In the absence of any further information, there are nine B1's, seven B2's, and twelve B3's, so we would guess B3, and expect to be right about 12 out of 28 times, giving us an error expectation of 16 out of 28, or about 57%. We will call this the *absolute error rate of B.* Now, however, suppose we are given a new data point, and are told its value for attribute A. How will we predict B, and what will our expected error rate be when conditioned on this knowledge? Well, 13/28 of the time the new point will be A1, and we should then guess B3, and expect to be right 7/13 of the time. Similarly, 7/28 of the time it will be A2, and we will guess B2, and be right 4/7 of the time, and 8/28 of the time it will be A3, we guess B3, and are right 5/8 of the time. Predictions of B conditioned on A, then, should be correct $(13/28)(7/13) + (7/28)(4/7) + (8/28)(5/8) \approx 57\%$ of the time, and the *error rate of B conditioned on A* is 43%. The *reduction in error* is $57 - 43$, and the *proportional reduction in error* is $(57 - 43)/53 \approx 26\%$. In comprehensible terms, we expect to be wrong about 26% fewer times if we know A. The proportional reduction in error when predicting A conditioned on B can be computed similarly. The absolute error rate of A is $(28 - 13)/28 \approx 54\%$, the error rate of A conditioned on B is $1 - [(9/28)(5/9) + (7/28)(4/7) + (12/28)(7/12)] \approx 43\%$, and the proportional reduction in error is $(54 - 43)/54 \approx 20\%$. Each of these proportional reductions

in error is a measure of how well knowledge of one attribute aids the prediction of the other. A symmetric measure of association can be obtained by simply averaging the two conditioned measures, giving the symmetric λ, of 23%.

Obviously, the more strongly two attributes are associated, the higher the value of λ, and *vice versa*. Some other properties of λ (Goodman and Kruskal, 1954) are:

- λ lies between 0 and 1, inclusive, except when the entire population lies in a single cell of the table, in which case it is indeterminate.
- λ is 1 if and only if all the population is in cells no two of which are in the same row or column.
- Independence is sufficient, but not necessary, for λ to equal 0.
- λ is unchanged by permutations of rows or columns.

Elsewhere we have found λ to be an excellent measure of qualitative association, in that it accords well with human intuitions and is much more "stable" than χ^2 (Chen, 1992). Using λ to calculate the association between cluster-numbers and categorical attribute values is faithful to the comprehensibility postulate: an attribute is a good description of a clustering if knowledge of the attribute helps predict cluster, and *vice versa*.

3.2 Integrating qualitative and quantitative data

The frequency table approach works well for categorical variables, but what about numeric variables? Nonmetric clustering takes a pragmatic approach to these: if we assume that the data are going to be adequately described by a clustering into a finite number of clusters, then there are really only a finite number of values of a numeric parameter to consider, one for each cluster. All other variations in a numeric parameter can be assumed to be due to variance within the clusters. Accordingly, we can divide up the range of a numeric parameter into discrete parts. We can do this nonmetrically by simply choosing quantile points, but a more flexible arrangement allows the "splits" between categorically different values to be selected by the algorithm as it runs. How this is accomplished is illustrated in Figure 1. Here we have marked two clusters with open and filled circles, and the categorical division of two dimensions into "high" and "low" values are shown by the dividing gray lines. The point marked with an "X" is troublesome, as it does not fit well with either of the two clusters, and keeps us from obtaining a λ value of 1.0 for this data set. We could move the vertical line to the right, to try to include X in one cluster, but that would raise more problems by the inclusion of some points from the other cluster. Similar problems occur if we try to raise the horizontal line.

The computer program Riffle will keep adjusting these split lines up and down to achieve better associations between cluster and numeric attribute. In

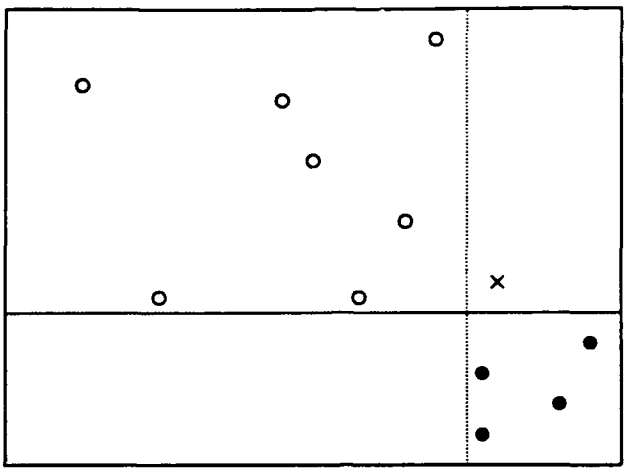


Figure 1: Twelve data points in two dimensions. Clusters are indicated by open and filled circles. Split values shown by gray lines. The point marked with an "X" cannot be included in either cluster by moving the split values without introducing further problems.

other words, what counts as "small" or "large" can be redefined by the algorithm as it investigates the data. At the same time, the algorithm is free to reassign the points themselves to different clusters. Both of these reinterpretations of the data are tried over and over, to maximize λ. The algorithm stops when it cannot improve the association between clusters and attributes any more, by any of its tricks.

This clustering methodology has a number of advantages over traditional clustering methods:

- It does not combine counts from dissimilar taxa by means of sums of squares, or other *ad hoc* mathematical techniques.
- It does not require transformations of the data, such as normalizing the variance.
- It works without modification on incomplete data sets. Since each attribute has its λ-association with the clustering evaluated independently, the fact that some points have some values for some attributes, and other points for other attributes, is irrelevant. Attributes are *not* directly combined.
- It can work without further assumptions on different data types (*e.g.*, numeric, categorical, species counts, presence/absence data, *etc.*).

- Significance of an attribute to the analysis is not dependent on the absolute size of its count. For instance, a taxon having a small total variance, such as rare taxa, can compete in importance with common taxa, and taxa with a large, random variance will not automatically be selected, to the exclusion of others.
- It provides an integral measure of "how good" the clustering is, *i.e.* whether the data set differs from a random collection of points, by means of the size of the λ values for each attribute.
- It can, in some cases, identify a subset of the attributes that serve as reliable indicators of the physical environment. In our research the indicator species selected by Riffle often proved to be more reliable than indicators based on a linear discriminant (Matthews et al., 1991a; Matthews et al., 1991b).

The major disadvantage of the Riffle program is that, in order to find a clustering of the data points with the desirable qualities listed above, a massive search through thousands of potential clustering candidates is made before settling on the "right" one. Even after this search, there is no guarantee that Riffle finds the optimal clustering, in the sense outlined above. However, in our research, Riffle does find an excellent clustering in a reasonable amount of time. For larger datasets, supercomputers and/or more heuristic searches may be required.

4 Association Analysis: a Significance Test from the Clustering

If the data analyzed have natural groups, such as treatment groups or sites, a significance test can be derived from the known groups and the generated clusters. Under the null hypothesis, clusters generated from the data will have no association with the known treatment groups. Thus, if the generated clusters closely match the treatment groups, with less than one or five percent probability under the null hypothesis, then a significant effect has been found. We have used nonmetric clustering and association analysis on a variety of multivariate experiments and find it to be comparable in sensitivity to many metric tests that make more assumptions about the underlying distributions of the data (Landis et al., forthcoming).

5 Implications for Ecological and Ecotoxicological Tests

The fact that nonmetric clustering and association analysis (NCAA) adheres to the comprehensibility postulate has numerous consequences for the analysis

of ecological data, and for policy. When establishing policy for mitigation or restraint, the ecologist is forced into the position of deciding what is "good" and what is "bad," or natural *vs.* unnatural, or pristine *vs.* polluted, or healthy *vs.* unhealthy. The development of various ecological indicators (diversity indices, indicator species, biomarkers, *etc.*) has proceeded by fits and starts, primarily because ecosystems are complex and rarely reproducible, and so a simple division into good and bad ecosystems is not feasible. Instead, each new system must be approached on its own terms, and ecological and toxicological experts must begin to understand it afresh and derive new concepts each time.

A computational induction from the data alone using ML techniques, on the other hand, has a number of advantages.

1. Machine learning is free from prejudice. Too often natural ecologists are forced to rely on traditional indicator species, or traditional measures of diversity, rather than taking a fresh look at each new system. Machine learning software does not remember the past.

2. Machine learning is adaptable. There is no need to establish policy based on a few preselected species, or on one mathematical technique. A variety of techniques, and all possible species, can be incorporated into a single ML tool which will sort through them and return with an objective picture of the ecosystem based on the most interesting species and the most informative tools.

3. Machine learning is interactive. Because the concepts derived by computational induction are faithful to the comprehensibility postulate, they can be examined by human experts. The machine is not a "black box" which must either be trusted implicitly or thrown out completely. Refinements in the ML algorithm can be visualized, based on experiments, and reincorporated into future generations of the ML computational tools.

4. Machine learning is not constrained like expert systems. Unlike expert systems, which attempt to encapsulate a particular human's expertise in a computer system, ML tools attempt to derive new expertise, new categories and concepts, derived from the data themselves. The only constraint on an ML system is the comprehensibility postulate, requiring that all new ideas be expressible in human terms. Beyond that, anything goes.

5. Machine learning is inexpensive. One of the primary motivations behind the surge of interest in expert systems was that a computer program represents a large initial investment, but a very small marginal cost subsequently, compared to professional consultation with a human expert. ML systems, once developed, are marketed like any other software, and can be duplicated and reused, in identical form, on any site.

Because of these advantages, we can recommend a new direction in ecotoxicological policy. There is a middle ground between reliance on completely

objective, simple, numerical cutoffs, on the one hand, and largely subjective, naked faith in consensus human judgement, on the other. Rather, policy must be made only after extensive interaction between human experts and their ML assistants. Without ML and the associated computational induction, the human expert cannot be sure that some important concepts not are being overlooked. The human's compromises and policies should only be made after the minimal step of consulting with an ML system. Such man-machine consultations must become part of policy, or else we are condemned to base judgements on only partial information, on oblique, narrow, and slanted views into the data. We therefore call for ecotoxicologists to review the large ML literature, and begin to establish standards for human-computer interactive analysis of ecological systems.

6 Future Work: Dynamic Ecosystem Change

While our system of nonmetric clustering and association analysis does well with a variety of environmental data, we are currently seeking a much-needed extension of our ideas. At present, each data set is treated statically, as an independent point in time. In reality, environmental systems are extremely sensitive to their history. What is needed is a conceptual description of ecological systems that pays particular attention to the dynamic nature of systems over time. On the one hand, time could simply be viewed as another measured attribute; however, it is obvious that this attribute holds a special place. Time series analysis, as it is currenly practiced, is almost entirely a univariate technique, primarily concerned with trends and cycles. What is required is a multivariate technique that makes sense of multivariate trends in patterns. One straightforward approach is to consider the state of a multivariate system as a multivariate vector, and the change over time as simply another vector connecting the state at one time with the state at another. In this view, we could define velocity, curvature, torsion, and a host of other vectors which would, in some sense, characterize the changes of the system over time. However, we must look instead for a description of change that does not violate the comprehensibility postulate. For a conceptual clustering, we must look for a *conceptual shift*, and have a concise notion of what this means. When we have decided the terms under which conceptual shifts are described, we can then build an ML tool that will assist us in our search for understanding. We believe that a conceptual shift in the character of a community or ecological system will be far more significant than any simple change in the numbers of species.

7 Conclusion

Machine learning promises to revolutionize the practice of environmental policy, by making the marriage of human and computer expertise a reality. We anticipate computerized "policy assistants" that will create an atmosphere of understanding and familiarity with the most difficult data. We have presented here, as an illustration, our own technique of nonmetric clustering and association analaysis, which we have used repeatedly in gaining deeper insights into ecological and toxicological data. All analysts who use only fixed methodologies, or only intuition, or both, in examining complex data, do so at their peril. The computer tools of machine learning present a new alternative to past practices, one which is at the same time more friendly and more objective, and one which will, sooner or later, be indispensible to our field.

References

Asimov, D. (1985). The grand tour. *SIAM Journal of Scientific and Statistical Computing*, 6:128–143.

Chen, C. (1992). The measurement of clustering tendency in machine learning. Master's thesis, Western Washington University, Bellingham Washington.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49:732–764.

Goodman, L. A. and Kruskal, W. H. (1959). Measures of association for cross classifications ii: further discussion and references. *Journal of the American Statistical Association*, 54:123–163.

Goodman, L. A. and Kruskal, W. H. (1963). Measures of association for cross classifications iii: approximate sampling theory. *Journal of the American Statistical Association*, 58:310–364.

Goodman, L. A. and Kruskal, W. H. (1972). Measures of association for cross classifications iv: simplification of asymptotic variances. *Journal of the American Statistical Association*, 67:415–421.

Huber, P. J. (1985). Projection pursuit. *Annals of Statistics*, 13:435–475.

Landis, W. G., Matthews, R. A., Markiewicz, A. J., and Matthews, G. B. (Forthcoming). Multivariate analysis of the impacts of the turbine fuel jp-4 in a microcosm test with implications for the evaluation of ecosystem dynamics and risk assessment. *Ecotoxicology*.

Matthews, G. and Hearne, J. (1991). Clustering without a metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(2):175–184.

Matthews, G. B., Matthews, R. A., and Hachmöller, B. (1991a). Mathematical analysis of temporal and spatial trends in the benthic macroinvertebrate communities of a small stream. *Canadian Journal of Fisheries and Aquatic Sciences*, 48(11):2184–2190.

Matthews, R. A., Matthews, G. B., and Ehinger, W. J. (1991b). Classification and ordination of limnological data: a comparison of analytical tools. *Ecological Modelling*, 53:167–187.

Michalski, R. S. (1983). A theory and methodology of inductive learning. *Machine Learning, An Artificial Intelligence Approach*, pages 83–134.

Michalski, R. S. and Stepp, R. E. (1983). Learning from observation: Conceptual clustering. *Machine Learning, An Artificial Intelligence Approach*, pages 331–363.

Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end games. In Michalski, R. S., Carbonell, J. G., and Mitchell, T. M., editors, *Machine Learning, An Artificial Intelligence Approach*, pages 463–482. Morgan Kaufmann, Los Altos, California.